000-1st_pages:Layout 1 18.05.2008 23:03 Page 1

Компьютерная лингвистика и интеллектуальные технологии

по материалам ежегодной международной конференции «Диалог» (2008)

Периодическое издание, выпуск 7 (14)

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2008)
Issue 7 (14)

000-1st_pages:Layout 1 18.05.2008 23:03 Page 2

УДК 80/81; 004
ББК 81.1
К63

Программный комитет конференции выражает искреннюю благодарность
Российскому фонду фундаментальных исследований
за финансовую поддержку, грант № _____

Редакционная коллегия:

А.Е. Кибрик (главный редактор),
В.И. Беликов, Б.В. Добров, Д.О. Добровольский,
Л.М. Захаров, И.М. Зацман, Л.Л. Иомдин,
И.М. Кобозева (ответственный секретарь),
Е.Б. Козеренко, М.А. Кронгауз, Н.И. Лауфер,
Н.В. Лукашевич, А.С. Нариньяни (зам. гл. редактора),
Г.С. Осипов, Н.В. Перцов, Т.В. Черниговская,
И.В. Сегалович, В.П. Селегей

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2008 г.). Вып. 7 (14).– М.: Изд-во РГГУ, 2008. 649 с.: ил.

Сборник включает 95 докладов международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2008», представляющих широкий спектр теоретических и прикладных исследований в области описания естественного языка, моделирования языковых процессов, создания практически применимых компьютерных лингвистических технологий.

Для специалистов в области теоретический и прикладной лингвистики и интеллектуальных технологий.

ISBN

© Институт проблем информатики РАН, 2008

© Редколлегия сборника «Компьютерная лингвистика и интеллектуальные технологии» (составитель), 2008 г.

000-1st_pages:Layout 1 18.05.2008 23:03 Page 3

Предисловие

Седьмой выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит материалы 14-й Международной конференции Диалог.

В программу конференции было отобрано 95 докладов, охватывающих все основные направления конференции:

- лексическая и грамматическая семантика и типология;
- корпусная лингвистика;
- информационный поиск и анализ документов;
- лексикография;
- машинный перевод;
- создание и использование тезаурусов и онтологий;
- речевые технологии;
- лингвистика коммуникации;
- инструментарии и специализированные базы данных для лингвистических исследований.

В полном соответствии со своим названием конференция является традиционным местом встречи и интенсивного обмена идеями между специалистами в области лингвистики, коммуникации, представления знаний, компьютерных технологий. Практической целью этого междисциплинарного общения является решение широкого спектра задач автоматической обработки естественного языка. Научная цель заключается в том, чтобы получить теоретические и языковые описания той степени полноты и эксплицитности, которая позволила бы решать вышеуказанные задачи. В этой ориентации инженерной практики на адекватные лингвистические и коммуникативные модели и состоит специфика Диалога, отличающая его от большинства конференций по компьютерной лингвистике.

Каждый год Программный Комитет предлагает какие-то актуальные проблемы в качестве доминант конференции. В этом году особое внимание уделено темам «Современная лексикография и лингвистика» и «Текстовые корпуса как объект и инструмент лингвистических исследований».

Темы эти непосредственно связаны, поскольку именно появление текстовых корпусов радикально изменило методику работы с языковым материалом как лингвистов, так и лексикографов. К сожалению, влияние лингвистической науки на лексикографическую практику до сих пор остается очень скромным, что негативно отражается и на возможности использования существующих лексикографических ресурсов в компьютерной лингвистике. Сближение прикладной лексикографии с теоретической лингвистикой является одной из важнейших целей Диалога. В сборнике представлены работы, отражающие новые подходы к языковому материалу и словарному содержанию, новые лексикографические объекты, новые типы словарей.

Создание и использование текстовых корпусов давно стало одной из самых актуальных тем на Диалоге. В этом году в корпусном направлении ярко представлена новая и очень важная область: корпуса звучащей речи. Помимо этого серьезное внимание уделено важнейшей проблеме лингвистической разметки текстов, решение которой позволяет гораздо более эффективно использовать корпуса как для лингвистических исследований, так и для машинного обучения.

Как всегда, корпусная тема включает в себя и проблемы использования Интернета как лингвистического ресурса.

Невозможно в кратком предисловии анонсировать всё, что представлено в этом сборнике. Тематика Диалога очень широка: мы рекомендуем сайт конференции www.dialog-21.ru всем, кому интересны проблемы компьютерной обработки естественного языка. На этом сайте можно ознакомиться и с условиями участия в конференции и публикации в этом ежегоднике. Там же представлены обширные электронные архивы Диалога, включая тексты всех сборников и доклады, не вошедшие в них.

Программный комитет Диалога
Редколлегия ежегодника «Компьютерная лингвистика и интеллектуальные технологии»

000-1st_pages:Layout 1 18.05.2008 23:03 Page 4

Организаторы

Ежегодная конференция Диалог проводится под патронажем Российского Фонда Фундаментальных Исследований при организационной поддержке компании ABBYY. Основными учредителями конференции являются:

Институт лингвистики РГГУ
Институт проблем информатики РАН
Институт проблем передачи информации РАН
Компания ABBYY
Компания Яndex
Рос НИИ искусственного интеллекта
Филологический факультет МГУ

При поддержке Российской ассоциации искусственного интеллекта

Международный программный комитет

Нариньяни Александр Семёнович, председатель	РосНИИ искусственного интеллекта
Буате Кристиан	Гренобльский университет
Богуславский Игорь Михайлович	Институт проблем передачи информации
Гельбух Александр Феликсович	Национальный политехнический институт (Мехико)
Зарецкая Елена Наумовна	Академия народного хозяйства при Правительстве РФ
Кибрик Александр Евгеньевич	Филологический факультет МГУ
Кобозева Ирина Михайловна	Филологический факультет МГУ
Козеренко Елена Борисовна	Институт проблем информатики РАН
Кронгауз Максим Анисимович	Институт лингвистики РГГУ
Лауфер Наталия Исаевна	Издательство «Альпина Бизнес Букс»
Мельчук Игорь Александрович	Монреальский университет
Ниренбург Сергей	Университет Нью-Мексико
Осипов Геннадий Семёнович	Институт программных систем РАН
Попов Эдуард Викторович	РосНИИ информационной техники и систем автоматиз. проектирования
Сегалович Илья Валентинович	Компания Яndex
Селегей Владимир Павлович	Компания ABBYY
Сулейманов Джавдет Шевкетович	Институт информатики КГУ
Флюр-Семёнова Вера	Компания SCIPER
Ыйм Халдур	Тартуский университет

Организационный комитет

Селегей Владимир Павлович, председатель	Компания ABBYY
Азарова Ирина Владимировна	Санкт-Петербургский государственный университет
Добров Борис Викторович	НИВЦ МГУ
Зацман Игорь Моисеевич	Институт проблем информатики РАН
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН
Лауфер Наталия Исаевна	Издательство «Альпина Бизнес Букс»
Лукашевич Наталья Валентиновна	НИВЦ МГУ
Перцов Николай Викторович	Авикомп Сервисез
Соколова Елена Григорьевна	РосНИИ искусственного интеллекта
Толдова Светлана Юрьевна	Филологический факультет МГУ
Шаров Сергей Анатольевич	РосНИИ искусственного интеллекта

Секретариат

Гайван Анна Александровна, секретарь оргкомитета, редактор сайта	Компания ABBYY
Бронникова Ольга Васильевна, координатор	Компания ABBYY

000-1st_pages:Layout 1 18.05.2008 23:03 Page 5

Рецензенты

Азарова Ирина Владимировна
Баранов Анатолий Николаевич
Баранов Юрий Владимирович
Беликов Владимир Иванович
Гельбух Александр Феликсович
Добров Борис Викторович
Добровольский Дмитрий Олегович
Ермаков Александр Евгеньевич
Зарецкая Елена Наумовна
Захаров Леонид Михайлович
Зацман Игорь Моисеевич
Иомдин Борис Леонидович
Иомдин Леонид Лейбович
Кибрик Андрей Александрович
Кибрик Александр Евгеньевич
Кобозева Ирина Михайловна
Козеренко Елена Борисовна
Крейдлин Григорий Ефимович
Кронгауз Максим Анисимович
Крылов Сергей Александрович
Лауфер Наталия Исаевна
Левонтина Ирина Борисовна
Лобанов Борис Мефодьевич
Лукашевич Наталья Валентиновна
Нариньяни Александр Семёнович
Осипов Геннадий Семёнович
Перцов Николай Викторович
Петрова Мария Андреевна
Плунгян Владимир Александрович
Подлесская Вера Исааковна
Рахилина Екатерина Владимировна
Савельев Василий Евгеньевич
Сегалович Илья Валентинович
Селегей Владимир Павлович
Семенова-Флюр Вера Эммануиловна
Сокирко Алексей Викторович
Соколова Елена Григорьевна
Тестелец Яков Георгиевич
Толдова Светлана Юрьевна
Туганбаев Диар Аскарович
Циммерлинг Антон Владимирович
Шаров Сергей Анатольевич
Янко Татьяна Евгеньевна
Янович Игорь Сергеевич

001-TOC:_Layout 1 18.05.2008 23:02 Page 6

Содержание

001-TOC:_Layout 1 18.05.2008 23:02 Page 7

001-TOC:_Layout 1 18.05.2008 23:02 Page 8

001-TOC:_Layout 1 18.05.2008 23:02 Page 9

001-TOC:_Layout 1 18.05.2008 23:02 Page 10

002-Azarova_Grebenkov_Lando:_Layout 1 18.05.2008 14:12 Page 11

ИСПОЛЬЗОВАНИЕ МАРКЕРОВ АКТАНТНЫХ ПОЗИЦИЙ ПРИ АНАЛИЗЕ ДЕЛОВЫХ ТЕКСТОВ ДЛЯ РАСШИРЕНИЯ ЛОГИЧЕСКОЙ СХЕМЫ ПРЕДМЕТНОЙ ОБЛАСТИ

THE CONTEXT SCHEMA OF PREDICATE ARGUMENTS FOR AUTOMATIC EXPANSION OF A DOMAIN ONTOLOGY

Азарова И.В. *(ivazarova@gmail.com),* ***Гребеньков А.С.*** *(shurix@grebenkov.ru),*
Ландо Т.М. *(tatiana.lando@gmail.com)*
Санкт-Петербургский государственный университет

В докладе представлено описание системы анализа деловых текстов Фактус для ограниченной предметной области с использованием синтактико-семантических шаблонов и маркеров актантных позиций. Обсуждается возможность расширения логической схемы заданной предметной области на основании анализа текста и создания таким образом «открытой понятийной системы».

1. Введение

На Кафедре математической лингвистики Санкт-Петербургского государственного университета был выполнен ряд исследований [1, 2] по автоматическому анализу текста с использованием имеющихся в распоряжении коллектива ресурсов – компьютерного тезауруса RussNet и формально-грамматического парсера Rus4IR. Была представлена пропозициональная модель семантического анализа текста [2], однако для решения практических задач необходимо прояснить ряд вопросов.

Одной из наиболее важных проблем является то, каким образом соотносится идеографическая структура тезауруса RussNet и логико-понятийная схема некоторой предметной области. RussNet в силу методики построения [3] фиксирует общую для различных функциональных стилей активную часть современного русского языка. В традиции построения wordnet-словарей существует представление о том, что терминологические «расширения» базовой структуры будут естественным образом присоединяться к имеющей основе, создавая более подробную понятийную классификацию [9], хотя и отмечалось, что семантическая интерпретация словосочетаний (особенно терминологических) требует особого подхода. Поскольку маловероятно исчисление всех комбинаций компонентов, образующих терминологические и нетерминологические словосочетания, необходимо понять, каким образом и в каком объеме следует предварительно фиксировать основные понятия предметной области, как их комбинировать и дополнять; каким образом будут соотносится понятия логической схемы предметной области и идентификаторы конкретных объектов (так называемые «именованные сущности»)?

Еще одним нерешенным вопросом является то, как представлять содержание текста: в пропозициональной структуре, предложенной ранее, или в какой-то иной форме [8, 5].

Для прояснения перечисленных вопросов была создана модельная система анализа деловых текстов Фактус, которая сама по себе имеет практическое значение, но при этом должна позволять естественно расширять как имеющуюся структуру понятий, так и вводить новые структуры. В качестве предметной области мы выбрали «назначения на должности управляющего персонала в компаниях», она довольно часто используется для иллюстрации возможности анализа текстов [4, 6] и имеет собственную информационную ценность.

2. Модель предметной области
2.1 Логическая схема предметной области

Для создания логико-понятийной схемы предметной области были исследованы тексты (25 текстов в среднем по 1 000 знаков) по данной тематике (CNEWS.ru и подобные). Поскольку тексты содержали разнообразные сведения, было отобрано «ядро», которое впоследствии может быть расширено, − это собственно данные о назначении руководителей высшего звена в компаниях различного типа. Другие сведения: о профильной продукции компаний, выпуске новых видов продукции, участии в конкурсах и различных проектах, а также биогра-

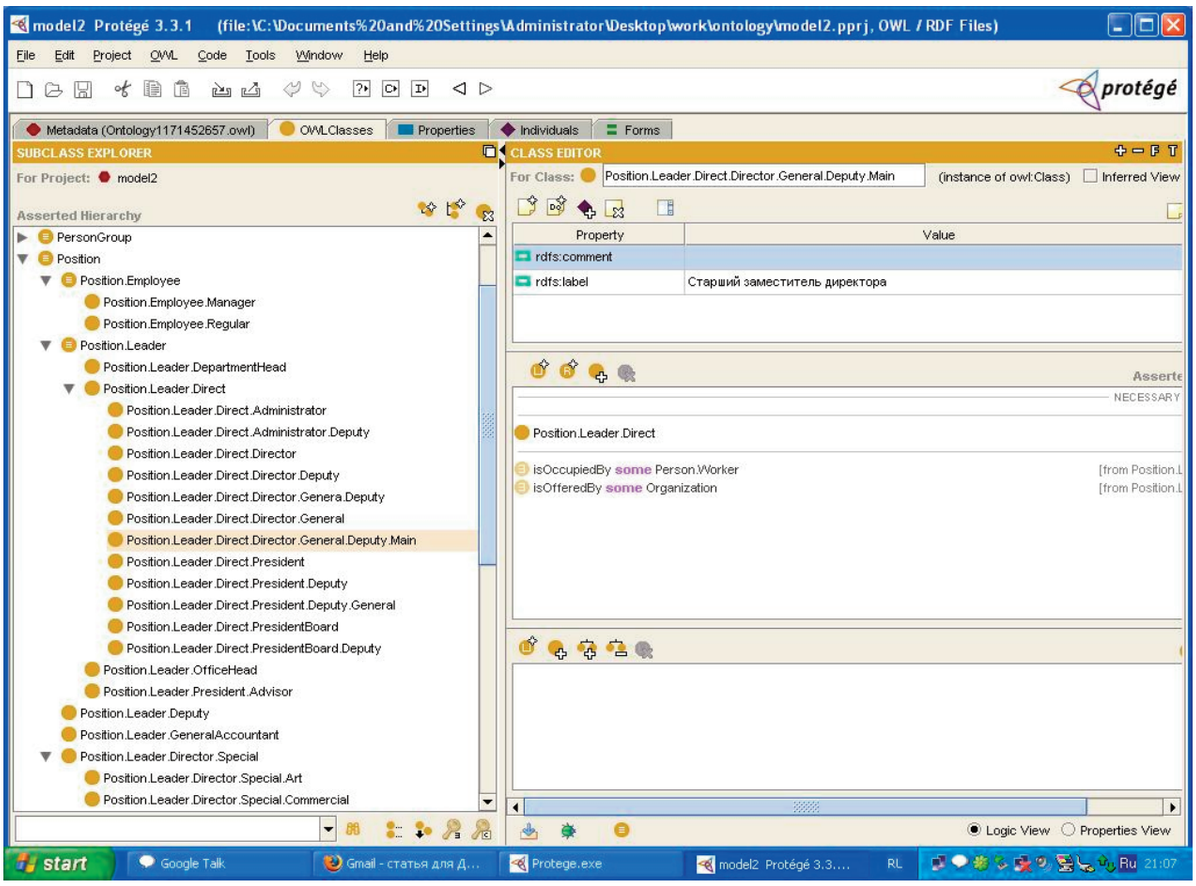
002-Azarova_Grebenkov_Lando:_Layout 1 18.05.2008 14:12 Page 12

фические сведения о руководителях – при необходимости будут использованы для расширения границ предметной области.

Мы учитывали, что возможны варианты онтологического представления предметной области, в которых могут быть акцентированы в зависимости от целей моделирования:

1. приближенность к объективной картине мира;
2. соответствие описания толковым словарям, энциклопедиям или другим ресурсам;
3. удобство обработки итоговой онтологии;
4. потенциальная возможность автоматически расширять онтологию.

Начальная классификация понятий предметной области была создана вручную, данные были представлены в формате OWL в редакторе Protégé 3.3.1. Обработка текстовых данных позволила зафиксировать так называемые «открытые» классы понятий с тем, чтобы далее использовать возможности автоматического дополнения логической схемы, или онтологии, предметной области. Помимо прочих соображений, учитывалась возможность относительно простого соотнесения со структурой wordnet-словарей.



В качестве основных классов были выбраны те, которые имеют референциальную соотнесенность: ***Компания***, ***Человек***, ***Группа людей*** (коллективные органы управления, например, *совет директоров*). Еще одна группа понятий ***Должность*** – руководящие должности в компаниях – на первый взгляд может быть представлена в виде отношения между членами классов ***Компания*** и ***Человек***. В частности, названия руководящих позиций (*директор*, *заместитель* и проч.) представлены в Принстонском WordNet в структуре семантического дерева ***Person*** *(individual, someone, mortal, human, soul)*, а в SUMO [10] – класс ***Position*** входит в класс ***SocialRole*** и является классом атрибутов, указывающих область ответственности в некоторой организации. Связь элементов класса ***Должность*** задается предикатом «занимать позицию» (***occupiesPosition***), который имеет 3 аргумента, члены классов ***Человек***, ***Должность***, ***Организация***.

(instance occupiesPosition TernaryPredicate)
(domain occupiesPosition 1 Human)
(domain occupiesPosition 2 Position)
(domain occupiesPosition 3 Organization)
(documentation occupiesPosition EnglishLanguage «(&%occupiesPosition ?PERSON ?POSITION ?ORG)

002-Azarova_Grebenkov_Lando:_Layout 1 18.05.2008 14:12 Page 13

В нашем исследовании онтологическое представление используется в первую очередь для эффективной и специфической обработки данных, поэтому интерпретация понятий должностных позиций в духе SUMO, как класса, больше подходит для фиксации ядра понятийной схемы системы Фактус.

При поверхностном рассмотрении может показаться, что между классами ***Человек***, ***Организация*** и ***Должность*** есть довольно четкое различие: первые два имеют трудно перечислимое множество «экземпляров», которые соотносятся с реальными объектами мира, а должности – закрытый класс понятий. Однако, в реальных текстах ситуация не так проста. И хотя имеется ограниченное количество названий высших руководителей в компании (*генеральный директор, президент, управляющий*), для многих высших и средних руководящих должностей в компаниях имеются собственные структуры, которые довольно сложно поддаются «выравниванию», например: *директор по безопасности* vs *директор по безопасности и режиму*, *директор по информационной политике и связям с общественностью* vs *директор по информационным технологиям* и проч. Таким образом, текстовые реализации понятий для этих трех классов имеют сходные черты: они представляют собой «открытые» списки, среди которых есть небольшое количество частотных экземпляров и ряд не поддающихся исчерпывающему перечислению именованных сущностей.

В данной работе мы опирались на удобство обработки понятийной модели и ее потенциальную расширяемость, что в отдельных случаях «навязывало» нам некоторые решения. Например, в структуре класса ***Должность*** есть подструктура заместителей (*директора, директора по маркетингу* и проч.). В Принстонском WordNet различного рода заместители помещены в качестве гипонимов понятия «помощник» {assistant, helper, help, supporter}, то есть отдельно от поддерева с вершиной «руководитель» {leader}. В нашей модели мы посчитали неразумным создавать параллельную структуру заместителей для руководителей различного рода и сочли возможным при необходимости подчинять элементам класса ***Должность «***заместителей», которые вводятся в логическую структуру в виде стандартного расширения. Например, для понятия *заместитель креативного директора* будет автоматически создаваться узел онтологии Position.Leader.Director.Special.Art.Deputy на основе имеющегося понятия *креативный директор* Position.Leader.Director.Special.Art. Текстовые варианты названий должностей включаются в модель предметной области в редакторе Protégé 3.3.1 в качестве метки (label), поскольку для идентификации узлов онтологии используется латиница.

2.2. Фиксация «именованных сущностей»

Отдельной проблемой при моделировании предметной области является работа с именованными сущностями (named entities), то есть с идентификаторами отдельных индивидов, компаний и некоторых должностей. Формат OWL, в котором строится модель, поддерживает включение «экземпляров» или «индивидов» в онтологию, хотя возможно также создание внешних словарей-расширений для перечисления таких данных.

На первый взгляд, оба подхода вполне равноправны. При автоматическом анализе текста имена собственные желательно обрабатывать специальным образом, поскольку только частотные (и, как правило, исконные) могут быть перечислены заранее, поэтому мы считаем, что включение частичных перечней именованных сущностей в онтологию будет некорректным. Мы храним их в отдельных словарях-реестрах, а для связи конкретного имени собственного с нужным классом достаточно ввести дополнительное поле, в котором будет содержаться ссылка на понятийный узел онтологии (об этом подробно будет сказано ниже).

Особую проблему представляет отождествление именованных сущностей: текстовые варианты ссылок на лица и компании различаются. И если варианты упоминания лица достаточно предсказуемы (например: *г-н Иванов, Петр Иванов, Петр Сергеевич Иванов, П. Иванов, П.С. Иванов*), то варьирование названий компаний носит более сложный характер (например: *Siemens, Siemens CT, Siemens Corporate, Siemens Corporate Technology, Сименс* и т.д.). Разные ссылки на лицо в пределах указанной вариативности в одном тексте объединяются, при этом в качестве основного варианта выбирается наиболее развернутое именование (как правило, это первое упоминание в тексте), для остальных вариантов названий лиц в реестре даются ссылки на основной вариант. Для вариантов названия компаний вводятся обязательные и факультативные области (будут описаны в следующем разделе), отождествление имен производится при совпадении названия в основной части. Автоматическое расширение реестров именованных сущностей возможно при учете маркеров принадлежности к некоторому классу сущностей (маркеров актантных позиций), которые будут описаны в следующем разделе.

3. Маркеры актантных позиций

Центральное место при автоматическом анализе текстов в системе Фактус занимает выявление реализации предиката ***occupiesPosition***, который связывает между собой понятия 3 классов онтологической схемы, при этом текстовые реализации понятийных классов занимают актантные позиции. Каждый из рассматриваемых классов понятий (***Человек***, ***Должность***, ***Организация***) имеет определенные характеристики текстовых реализа-

002-Azarova_Grebenkov_Lando:_Layout 1 18.05.2008 14:12 Page 14

ций, на основании которых можно отнести то или иное слово или словосочетание к выражению понятия определенного класса. Совокупность таких признаков будем называть маркерами актантных классов, выделим среди них следующие типы (возможны также их комбинации).

1. Лексические маркеры

В качестве лексических маркеров выступают слова и выражения, которые явным образом задают «роль» актантной позиции, например, *господин, г-н, госпожа* и проч. для ***Человек***, *компания, в компании, концерн*, и проч. для ***Организация***, *на должность, на позицию, в должности, пост* и проч. для ***Должность***. В дальнейшем будем обозначать их как ролевые лексические маркеры. Кроме того, лексическими маркерами являются представленные в словаре компоненты текстовых реализаций для каждого понятийного класса, которые имеют собственную линейную структуру, включающую обязательные и факультативные компоненты. Позиция ***Человек*** задается указанием имени, отчества и фамилии, последний компонент является обязательным. ***Организация*** включает характеристику типа собственности компании и названия профиля организации, которые чаще выражены аббревиатурами (например, *ООО, ЗАО, КБ, ГПИ* и проч.), однако встречаются и в виде полных словосочетаний (*строительная компания, инвестиционный банк, коммерческий банк, акционерная аудиторская фирма* и т. п.), собственно названия – имени собственного в развернутом или сокращенном виде (*Ситроникс, Скай Линк* и т. п.) и указания территориальной принадлежности организации (*Уральский, Санкт-Петербургский, американский, в Дрездене* и проч.). Обязательным компонентом является имя собственное − название организации. Позиция ***Должность*** включает существительные, которые формируют ядро словосочетаний (*директор, управляющий, президент, заместитель* и т. п.) и зависимых слов – всевозможных атрибутов, представленных в виде определений и предложных конструкций (*старший вице-президент по продажам, директор по продаже профессиональных услуг* и проч.). Факультативные компоненты, оформляющие текстовое именование для класса ***Организация***, также используются как ролевые лексические маркеры: в первую очередь, это аббревиатуры типа *ООО, ЗАО* и т. д.

2. Пунктуационные маркеры

Четким пунктуационным маркером имени собственного компании являются кавычки, причем они выделяют как обязательный компонент названия, так и тип профиля/ владения организации, например, *ЗАО «Петер-Сервис», ОАО «Концерн «Ситроникс»*. Другим пунктуационным маркером являются скобки, которые используются для введения иноязычного названия организации или сокращенного обозначения, которое расценивается авторами текста как не вполне общепринятое, но в дальнейшем аббревиатура свободно используется в тексте: *ОАО «Концерн Научный Центр» (КНЦ), ОАО «Уралсвязьинформ» (URSI)*.

3. Графематические маркеры

Названия компаний и имена указываются в текстах в латинском написании чаще при пояснении в скобках и реже в основном тексте: *Алекс Адамопулос (Alex Adamopulos), компания Direct Tech Inc., Siemens AG*. Маркером имен собственных при обозначении лиц и организаций является написание с заглавной буквы, однако при этом необходимо отделять их от маркера начала предложения. Использование одних прописных в слове маркирует как аббревиатуры, так и собственные имена компаний: *ОАО МЕДУЗА*.

4. Маркер новизны

В том случае когда у предиката ***occupiesPosition*** есть незаполненные позиции, то графематически отмеченные слова и словосочетания (предположительно, это имена собственные), которых нет в словаре, получают дополнительный вес в качестве заполнителей актантных позиций. Для определения типа позиции большое значение имеют факультативные ролевые лексические маркеры (например, *ЗАО БМК-АВТО*). Такие текстовые фрагменты будут автоматически подключаться к словарю.

5. Синтаксические маркеры

В качестве частотной текстовой реализации предиката ***occupiesPosition*** выступают глаголы, которые предопределяют «конфигурации» актантных позиций (рамки валентностей) как в плане взаимного расположения актантов, так и способа морфо-семантического способа их оформления. Например, для пассивного залога глагола «назначить» (обычно в форме прошедшего времени *был назначен*) типичная конфигурация включает словосочетание-подлежащее, которое задает позицию ***Человек***, и дополнение в форме творительного падежа, обозначающее ***Должность***, при котором в форме предложной конструкции или генитива указывается ***Организация***: *Иван Петров был назначен директором ООО «Бригада»* или *Иван Петров был назначен директором в ООО «Бригада»*.

Словосочетания, заполняющие позицию ***Должность***, могут представлять ряды (*директор по информационной политике и связям с общественностью*), вариативность которых довольно велика, мы используем декомпозицию таких конструкций. Аналогичным образом декомпозиция применяется при анализе сложных описаний позиции ***Организация***, включающих факультативные компоненты в форме свободных словосочетаний (*ОАО Акционерная производственно-ювелирная компания «Золото Якутии»*).

002-Azarova_Grebenkov_Lando:_Layout 1 18.05.2008 14:12 Page 15

4. Процедура анализа делового текста предметной области

При анализе текста в системе Фактус моделируется «заинтересованный взгляд», когда читатель пропускает неважное и выбирает то, что ему интересно. То есть это ситуация неполного семантического анализа текста.

Целевая семантическая структура – извлечение тернарных предикатов ***occupiesPosition*** (назначение и увольнение руководящих сотрудников в компаниях) с заполненными позициями. Следовательно, основной принцип процедуры анализа состоит в том, чтобы вычленять и анализировать только релевантные отрывки текста, пропуская все, не относящееся к трехкомпонентной схеме. Необходимо смоделировать взгляд читателя, который выхватывает из текста определенные маркеры для понимания текста, учитывая то, что на вход анализатора подаются неподготовленные тексты (например, ленты новостей), которые могут не содержать интересующие нас факты, этот подход − пропускать все лишнее − является экономичным и оправданным, так как в тексте из нескольких абзацев может быть только одно предложение, содержащее интересующий нас факт. Таким образом, система работает как фильтр, реагирующий на определенные «ключи».

В соответствие с этой концепцией можно выделить несколько этапов анализа.

1. Предварительная обработка текста: фрагментация текста на предложения и словоформы, определение различных графематических маркеров.

2. Морфологический и словарный анализ, определение стандартных грамматических показателей; за счет контекстных связей частично разрешается грамматическая омонимия.

3. Синтактико-семантический анализ, результатом которого является выделение трехкомпонентных фактов.

Первые два этапа стандартны для решения любой лингвистической задачи, так что опишем подробнее этап синтактико-семантического анализа. В настоящее время единицей обработки является предложение, в дальнейшем планируется работа с абзацами и текстами. Основой этого этапа анализа является использование семантико-синтаксических шаблонов (аналогов рамок валентностей), в которых задаются схемы реализаций фактов. На сегодняшний день подготовлено порядка 100 подобных шаблонов. Структурно шаблоны состоят из следующих элементов:

1. Ключевые слова, или способы выражения предиката ***occupiesPosition***, которые выражены лексически: *назначить* (активная или квазипассивная конструкция), *уйти*, *перейти* и т. д.

2. Связанные с конкретной формой предиката конфигурации актантов, то есть синтаксических маркеров.

3. Шаблоны декомпозиции для комплексных описаний компонентов ***Человек***, ***Должность***, ***Организация***.

Предложение сравнивается с шаблонами, описывающими искомые факты, вначале проверяется наличие ключевых предикатов, затем − конфигураций актантов. Вычисляется мера соответствия предложения некоторому шаблону, поскольку оно может соответствовать нескольким шаблонам одновременно. В результате выбирается шаблон с наибольшей степенью соответствия, по которому в предложении вычленяются текстовые реализации компонентов ***Человек***, ***Должность***, ***Организация.***

Например, шаблон, описывающий предложение *Вася Пупкин был уволен с поста директора фирмы РиК* в нашей нотации будет выглядеть так

(1) person:nom VERB:уволить:pass PREP:с post:gen org:gen,

где VERB и PREP являются терминальными элементами,

person, post и org задают нетерминальные структуры реализаций актантных позиций ***Человек***, ***Должность*** и ***Организация***.

Шаблоны в отличие от контекстно-свободных грамматик не задают жесткого порядка слов, поэтому указанному шаблону будут соответствовать и другие предложения *С поста директора фирмы РиК был уволен Вася Пупкин* и *С поста директора фирмы РиК Вася Пупкин был уволен*.

Одной из проблем использования шаблонов является создание полных перечней возможных синтаксических структур. Решением является автоматизированное пополнение набора шаблонов. Если в тексте встречается предложение, частично реализующее некий шаблон, при этом лишь частично определены компоненты факта, то есть основание предполагать, что в этом предложении скрыт шаблон, не внесенный в список. В таком случае это предложение попадает на проверку лингвисту.

Текстовые фрагменты, не зафиксированные в онтологии и словарях-реестрах именованных сущностей, которые выделены с использованием актантных маркеров, то есть получившие «подтверждение» их принадлежности к классу, автоматически добавляются, расширяют онтологию (перечни экземпляров).

002-Azarova_Grebenkov_Lando:_Layout 1 18.05.2008 14:12 Page 16

5. Выводы и перспективы исследования

В результате анализа модельной системы Фактус были получены некоторые ответы на поставленные в данной работе вопросы. Несмотря на то, что методика построения RussNet и онтологии предметной области Фактус опирается на анализ текстов, между структурами наблюдается довольно значительные отличия, которые, вероятнее всего, связаны с различием терминологического и нетерминологического употребления слов в деловом тексте и «усредненном», не имеющем четкой функциональной направленности тексте.

В отличие от чистых структур автоматического построения онтологий, в нашей системе реализован гибридный подход: первоначально ручная онтология расширяется автоматически, причем получаются плоские структуры. Также предполагается использовать синтаксические шаблоны совместно с порождающими грамматиками для описания структур компонентов. Мы надеемся, что этот подход позволит добиться значительного улучшения в поиске компонентов информации понятийной области.

Пропозициональная структура в качестве семантического представления имеет в большой степени теоретический характер, поскольку требует заполнения аргументных позиций, не всегда имеющих явного определения в тексте. Таким образом, она мало подходит для условий неполноты данных, особенно в ситуации неполной обработки текста. В такой ситуации более предпочтительной является опора на текстовое выражение информации – на то, что явно выражено в нем, и на формализованную систему фиксации информации, больше напоминая заполнение заранее заданных «слотов» − фрагментов информационных структур.

В настоящее время тестируется прототип системы, которая по завершению разработки будет включать в себя сразу несколько лингвистических механизмов анализа текста. Система проектируется таким образом, чтобы автономно обрабатывать большое количество реальных текстов деловой тематики.

В ближайшем будущем планируется соединить шаблонное представление структуры предложения или абзаца и порождающее описание локальных синтаксических групп (словосочетаний), для которых используется формализм порождающей грамматики AGFL [8]. Будет доработана схема объединения компонентов информации об одном факте или ряде фактов, которые представлены в двух или нескольких предложениях текста.

Список литературы

1. Азарова И.В., Иванов В. Л. Овчинникова Е. А. Использование схемы наследования рамок валентностей в тезаурусе RussNet для автоматического анализа текста. Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006». М., 2006. С. 18−25.

2. Азарова И.В., Овчинникова Е.А. Семантическая интерпретация именных конструкций по корпусу русских текстов Труды Международной конференции «Корпусная лингвистика – 2006». СПб, 2006. С. 25−33.

3. Азарова И.В., Синопальникова А.А. Использование статистико-комбинаторных свойств корпуса современных текстов для формирования структуры компьютерного тезауруса RussNet // Труды международной конференции «Корпусная лингвистика 2004». 11−14 октября 2004 г. СПб., 2004. С. 5−15.

4. Ермаков А.Е. Поиск фактов в тексте // Мир ПК. 2005. N 2.

5. Железняков М.М., Невлева Т.Н., Новицкая И.М., Смирнова Л.Н., Цейтин Г.С. Опыт построения модели типа «текст -> действительность» с использованием ассоциативных сетей // Машинный фонд русского языка: предпроектные исследования, Институт русского языка АН СССР, М., 1988, 140−167.

6. Киселев С.Л., Ермаков А.Е., Плешко В.В. Поиск фактов в тексте естественного языка на основе сетевых описаний // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’2004. М., 2004.

7. Azarova I.V., Ovchinnikova E. A., Ivanov V., Sinopalnikova A. A. RussNet as a Semantic Component of the Text Analyser for Russian. Proceedings of the Third International WordNet Conference. Brno, 2005. P. 19−28.

8. Koster C. H.A., Transducing Text to Multiword Units // Workshop on MultiWord Units MEMURA at the fourth International Conference on Language Resources and Evaluation, LREC-2004. Lisbon, Portugal, May 2004.

9. Harabagiu S.M., Moldovan D.I. Knowledge Processing on the Extended WordNet // WordNet: An Electronic Lexical Database / Ch. Fellbaum (ed.). MIT Press, 1998. P. 379–405.

10. Suggested Upper Merged Ontology (SUMO) // URL: http://www.ontologyportal.org/

003-Apresjan_V:_Layout 1 18.05.2008 13:59 Page 17

РУССКИЕ И АНГЛИЙСКИЕ ЭМОЦИОНАЛЬНЫЕ КОНЦЕПТЫ
RUSSIAN AND ENGLISH EMOTIONAL CONCEPTS[1]

Апресян В.Ю.*.(valentina.apresjan@gmail.com)*
Институт русского языка им. В.В.Виноградова РАН

В современной этнолингвистике одной из центральных является мысль о том, что различия в языке отражают различия в национальной ментальности. Соответственно, основная масса современных этнолингвистических работ, в том числе на русском материале, сосредоточена на поиске таких различий и, в частности, на сравнении «ключевых слов» разных языков (понятие, введенное А. Вежбицкой) – слов, которые являются особенно важными выразителями главных ментальных особенностей носителей того или иного языка (классические примеры А. Вежбицкой – это *душа, тоска, судьба*).

Признавая всю ценность этого подхода, в настоящей работе мы хотели бы, сосредоточиться не только на концептуальных различиях, но и на сходствах между языками. Работа выполнена на материале эмоциональных концептов в русском и английском языках. Наряду с использованием ставших классическими подходов к описанию эмоций в языке – прототипического подхода Л.Иорданской и А.Вежбицкой, принятого Московской семантической школой, а также метафорического подхода Дж.Лакова и З.Кевечеса, положивших начало когнитивистской школе, данная работа предлагает нечто новое. Целью является не сравнение отдельных слов или концептов, а сравнение целых концептуальных полей и построение широких семантических типологий. Рассматривается 11 групп (или кластеров) эмоций – ‘страх’, ‘гнев’, ‘отвращение’, ‘грусть’, ‘радость’, ‘стыд’, ‘жалость’, ‘обида’, ‘’, ‘гордость’, ‘ревность/зависть’, ‘благодарность’. В каждой группе анализируется основная масса лингвистических средств, выражающих разные стороны эмоции – во-первых, весь спектр синонимов в разных частях речи, представляющих разные типы и оттенки эмоции (например, для кластера ‘СТРАХ’ - *бояться, пугаться, трусить, страшиться, опасаться, страх, боязнь* и пр. для русского; *to be afraid, to be scared, to fear, fear, dread* и пр. для английского); во-вторых, весь спектр частеречно разнородных средств, представляющих разные аспекты эмоции – каузацию (*пугать, страшный*), эмоциональное состояние (*грустно, страшно*), поведение, мотивированное эмоцией (*ужасаться, жалеть*), физиологические реакции на эмоцию (*побелеть, побагроветь, похолодеть*); поведенческие реакции на эмоцию (*убежать в ужасе, ударить кулаком по столу от досады*). Проанализированный материал дает возможность установить как сходства, так и различия в концептуализации эмоций в русском и английском языках, а также в семантическом устройстве этих полей. Обнаруживается, что при бесспорном наличии многочисленных семантических контрастов между отдельными словами в русском и английском, а также частом отсутствии безусловного переводного эквивалента, устройство эмоциональных кластеров в целом у многих эмоций сильно пересекается. Например, и русский, и английский языки выделили следующие подвиды страха: общий, нейтральный страх (*бояться, страх, to fear/ to be afraid*), кратковременный биологический страх (*пугаться, to be scared, to be frightened*); сильный страх перед непосредственно угрожающим, масштабным и неизвестным (*ужас, terror*); сильный страх перед неизбежным и отдаленным во времени (*страшиться, to dread*); рациональный страх (*опасаться, to be apprehensive*); благоговейный страх перед высшими силами (*трепетать, to be awed*) и пр. При этом границы подвидов эмоций не обязательно совпадают с лексическими: так, *to be scared* может выражать и кратковременный “биологический” страх (*I got scared*), и нейтральный страх-отношение (*I’m scared of dogs*). Представляется, что сравнение не отдельных слов, а целых полей позволяет составить более объективное представление о языковой концептуализации каких-то явлений, в том числе эмоций, и избежать неверного отождествления отсутствия и наличия в языке каких-то слов – в частности, точных переводных эквивалентов словам другого языка, с отсутствием и наличием каких-то ментальных, когнитивных и эмоциональных особенностей у носителей этого языка.

The idea that languages both reflect and shape their speakers’ mentalities, goes back as far as Humboldt’s work and the famous Sapir and Whorf’s hypothesis. While SWH was empirically and theoretically challenged by the proponents of linguistic innateness and universality, it has seen a comeback in the works of Wierzbicka [Wierzbicka 1990, 1991, 1992, 1999] and the entire NSM school. Though SWH is hardly popular nowadays in its strong version, the weaker claim that language reflects mentality is very much a part of today’s ethnolinguistic discourse. It is strongly featured in Wierzbicka’s line of research, particularly in the notion of *key words* – “words which are particularly important and revealing in a given culture”, e.g. Russian *душа* ‘soul’, *тоска* ‘yearning’ and *судьба*

[1] This paper was written with the financial support of the grant for senior regional fellows 2007-2008, of the Davis Center for Russian and Eurasian Studies, Harvard, as well as partial financial support of the following grants: RF President grant for the leading scientific schools NSH-56.11.2006.6, RHSF N06-04-00289a for «Developing a word list and samples for an active dictionary of Russian», RHSF 07-04-00-202a for “System-forming meanings of Russian Language”, grant of the Program for Fundamental research of the Department of Humanities of the Russian Academy of Sciences «Russian culture in world history».

003-Apresjan_V:_Layout 1 18.05.2008 13:59 Page 18

'fate' [Wierzbicka 1990, 1997:15, 55-84].

In the last decade, the search for culture-specific vocabulary, including Russian data, has thrived, and researchers have added a few items to the list of Russian key words, among them *авось* 'perhaps with luck', *совесть* 'conscience', *жалость* 'pity', *истина* 'the Truth, gospel-truth', *друг* 'friend', *воля* 'unrestrained physical freedom', *смирение* 'humility' [Wierzbicka 1997:55-84, Bulygina & Shmelev 1997:481-495, Levontina & Zalizniak 2001: 306-309], *маяться* 'to hang about for a long time yearning', *томиться* 'to yearn, to languish', *неприкаянность'* 'the state of not knowing what to do and where to be, and feeling bad because of that', *позор* 'disgrace' [Shmelev 2002:404-410], *обида* 'offence, hurt feelings' [Levontina & Zalizniak 2001: 306-309], *собраться* 'to get around to doing something', *заодно* 'along with, while one is at it' [Shmelev 2002:300, 406-407], and others, which supposedly reflect various aspects of the "Russian soul".

While the contrastive ethnolinguistic framework has been extremely efficient in dispensing with Anglo-centricity in the description of language, as well as producing vastly impressive empirical results, it has also raised some general methodological questions. It is undeniably true that all cultures are unique and that these peculiarities are to some extent reflected in language. However, it is not entirely clear to what extent language can serve as an objective mirror of culture – in other words, to what extent it is possible to draw inferences about a culture or national mentality based on linguistic facts, such as a presence of a certain untranslatable word in the vocabulary of a language. Direct un-translatability is a regular phenomenon across languages, yet it in itself cannot serve as an indication that speakers of these languages do indeed view the world differently.

The notion of linguistic salience which is fundamental to the contrastive ethnolinguistic framework is not entirely transparent either: the frequency of a word in a language as compared to the frequency of its translation in another language is not necessarily a proof of greater salience of a respective concept, as it is often the case that a concept expressed by one polysemous word or a word with a more general meaning in one language is expressed by several words in another language, as is the case, for example, with the Russian word *душа* 'lit. soul', which corresponds to the English *soul* and *heart*.

Other linguistic factors that are often considered as an indication of a word's salience, such as, for example, the number of its derivatives, are not always reliable criteria either, as languages are bound to vary in this respect based on the variation in the richness of their grammatical and lexical systems, irrespectively of how central or marginal the concept in question might be.

This paper proposes a framework for a semantic typology of emotion concepts in language, which considers both their similarities and differences. The framework incorporates the existing linguistic accounts of emotions, i.e. the scenario-based semantic approach of [Iordanskaja 1972, 1984] and Wierzbicka [1999], also adopted by some psychologists [cf. Shweder 2004], as well the conceptual metaphor-based approach of the cognitive semantics [Lakoff & Jonhson, Kövecses 1990, Emanatian 1995, Kövecses 2000]. The novelty of the proposed approach is in its scale: rather than comparing individual parts of the system, such as separate emotion terms, entire systems in the form of "emotion clusters" are juxtaposed. Each emotion cluster is represented by many "members", e.g., *anger, fury, wrath, irritation*, and other expressions for 'ANGER' cluster in English, whose meanings together form the "conceptual map" of 'ANGER' system in that language. By cross-linguistic "superimposing " of the conceptual maps of emotions in different languages, one is able to capture a holistic picture of the emotional universe, where both the universality and the differences displayed across languages can be clearly seen. For example, both English and Russian have explicitly incorporated the following types of 'ANGER' in their conceptual systems: 'justified' anger (*indignation, негодование*), 'strong uncontrollable anger' (*rage, бешенство*), 'mean anger' (*venom, злоба*), 'nerve-wrecking anger' (*irritation, раздражение*), but English has also a type of 'helpless disappointed anger' (*frustration*), which Russian lacks.

Cluster comparison involves comparing multiple emotion terms belonging to different parts of speech and all other related linguistic items, denoting causation of emotions, emotional behaviors, physiological manifestations of emotions and other aspects of emotions.

Overall, 11 emotion clusters in English and Russian have been analyzed, including the five emotions considered "basic" in most psychological and physiological studies - 'FEAR', 'ANGER', 'SADNESS', 'DISGUST', 'JOY' (basic emotions), 'SHAME', 'OFFENCE', 'PITY', 'PRIDE', 'ENVY/JEALOUSY', 'GRATITUDE'.

Emotion clusters in English and Russian: general tendencies.

If one looks at emotion clusters in Russian and English, i.e., at an entire range of emotions within a certain group, one would find a lot of similarity in how the fields are organized logically and conceptually. While the precise linguistic expression can be and is different, a very similar range of emotion gradations is found in the two languages.

Moreover, emotion metaphors turn out to overlap to a large extent as well, which is hardly surprising given that a lot of emotion metaphors are biologically rooted and based on physiologically conditioned responses to various stimuli. Well-known examples of such biologically-based linguistic metaphors are FEAR IS COLD and ANGER IS HEAT metaphors (e.g., *to freeze with terror, to boil with anger*); cf. [Lakoff & Johnson 1980, Apresjan & Apresjan 1993, Kövecses 2000, on the metaphorical conceptualization of these emotions, Ekman 1984 on their physiological manifestations]. Recent neuropsychological research allows to expand this list, as it suggests that feeling "hurt" and feeling empathy indeed activate pain centers in the brain [McDonald & Leary 2005, Gallese 2001, Singer et al. 2004], thus providing a biological explanation for the widely spread OFFENCE IS PAIN and PITY IS PAIN metaphors (cf. *to be hurt, to be wounded, to be injured* as expressions of 'offended' feeling and *щемящая жалость* 'piercing pity', *больно за кого-то* 'to feel pained for somebody' as expressions of 'pity'). Likewise, [Calder et al. 2001] suggest that physical distaste, manifested in actual nausea, and moral repugnance that does not involve actual nausea symptoms, are nevertheless neurologically very much the same in that they activate the same neural pathways, thus proving a biological basis for another widely-spread metaphoric mapping, DISGUST IS FEELING SICK.

The following common tendencies in the organization of emotion clusters have been found:

First of all, both languages show a considerably larger number of emotion terms that denote unpleasant emotions, which is not surprising either biologically or linguistically. Among the biologically and psychologically defined "basic" emotions which presum-

003-Apresjan_V:_Layout 1 18.05.2008 13:59 Page 19

ably carry a survival value, there are four unpleasant ones (*fear, anger, sadness, disgust*) and only one pleasant (*joy*). Although there are more pleasant emotions among the non-basic ones (*pride, gratitude*), unpleasant on the whole outweighs the pleasant (*shame, pity*, *jealousy, envy, resentment*). Language, too, is well-known for marking negative and abnormal over positive or normal.

For each of the emotions examined, there are certain gradations within the cluster. Usually, there is a neutral term which denotes a natural degree of emotion in relation to the stimulus: *fear, anger, disgust, shame, joy, sadness, pride, pity,* even *jealousy* and their Russian correlates, that do not bear any negative or positive evaluation on the part of the speaker. A degree of emotion which is excessive in regard to the stimulus and therefore inappropriate, or a behavior driven by an excessively strong emotion, or being overly prone to experiencing a certain emotion is usually marked negatively: *cowardice* (behavior caused by one's inability to master fear)*, grumpy, irascible* (too prone to anger)*, uptight* (too prone to embarrassment)*, despondent* (too sad for too long a time)*, bleeding-heart* (too prone to pity), etc. Besides, within some clusters, there are terms denoting clinical conditions, marked by prolonged and unmotivated experiencing of a certain emotion: *phobia, fright*; *depression*; *aversion*.

Some emotions are expected to occur in appropriate circumstances and their absence is viewed as a deviation from the social or ethical norm, e.g., *pity, shame, gratitude*: thus, *pitiless, shameless, ungrateful* (*безжалостный, бесстыжий, бесстыдный, неблагодарный*) are negative terms marking the inability to experience a naturally expected emotion. Interestingly, *fear* is viewed as an expected response to danger, and its absence (*fearless, бесстрашный*) as a deviation from the norm, though in this case the norm is biological, and the deviation from it is not only socially acceptable, but even positively evaluated.

Emotions that are either socially expected or otherwise desirable responses can occur as a result of conscious stimulation; cf. *to cheer up, to shame, to move* or *веселить, стыдить, разжалобить*, which denote a controlled intentional action with the purpose of inducing the respective emotion.

Emotions which are not desirable because they are unpleasant for the experiencer or for their object or have no ethical value, usually occur as responses to unintentional stimuli, since nobody wants to induce them on purpose; cf. *to disgust, to sadden, to anger, to irritate*, *печалить,сердить,раздражать* which refer to unintentional behaviors or even events: *You disgust me, Her illness saddened him, He was angered by the result of the election, This noise irritates me*, but not **Перестань вызывать у меня отвращение* 'Stop disgusting me', *??Не печаль меня* 'Don't sadden me'. The idea of unintentional causation is also expressed by adjectives in both languages: *creepy, sad, scary*, *противный, печальный, страшный*, etc.

Emotions whose primary object is another person, can often be directed at self as well, but not if they are strong, uncontrollable, involve obligatory behavioral manifestations or too much of an alienation from the object; thus, one can have *self-pity, self-contempt, self-disgust*, *be angry at oneself*, or *испытывать к себе жалость <презрение, отвращение>, сердиться на себя*, but one cannot have **self-loathing, *rage at oneself, *have compassion for oneself*, **be offended by oneself* or **испытывать к себе сострадание <*гадливость>,*приводить себя в бешенство*..Likewise, the object of such emotions as *gratitude, envy, jealousy, благодарность,ревность,зависть* is always another person, not self, as they involve a great degree of alienation between the experiencer and the object.

Strong emotions can be manifested in either biological or near-biological reactions or in uncontrolled behaviors: cf. *to shake with fear, to choke with rage, to gasp with anger, to vomit with disgust, to cry with sorrow/pity, to laugh with joy, трястись от страха, задыхаться от ярости,тошнит от отвращения, плакать от обиды /жалости,смеяться от радости* (biological reactions) and *to flee in terror/panic, to strike in rage, в ужасе убежать, в гневе ударить* (uncontrolled behaviors), but not **to flee in apprehension, *to hit in annoyance, *убежать в опасении, *ударить в досаде*. Emotions which involve uncontrolled behaviors usually also involve biological reactions (*fury, terror*), but the reverse is not true (*joy, pity*) [cf. Mel'čuk & Wanner 1996 on the linguistic connections between an emotion's strength and controllability and its likely manifestations].

Strong emotions which do not deprive their experiencer of the ability to reason, such as *compassion, gratitude, envy* can drive him (her) to a controlled behavioral response: *to help smb. out of compassion, hire smb. out of gratitude, badmouth smb. out of envy*, *помочь кому-то из жалости, взять кого-то на работу из благодарности,оклеветать кого-то из зависти*, but not **flee out of panic, *kill out of fury*.

Along with these similarities, there are also a number of natural differences in the cluster organization in the two languages, with a somewhat different distribution of Russian-to-English discrepancies than it was previously thought. The existing "mismatches" do not necessarily reflect a fundamental difference in the emotional worlds and experiences in the speakers of Russian and English, but can be, to a great extent, accounted for linguistically. The following sources of cross-linguistic disparity in the area of emotion can be tentatively formulated:

The first source is different mapping of linguistic terms onto the same conceptual field, which is a phenomenon naturally found in all domains of language, not only in the sphere of emotion terms. It happens when a specific configuration of meanings is expressed by a single word in language X and by two or more words in language Y or even by some parts of their respective meanings (e.g., the word *horror* embraces the meanings of the words *ужас и отвращение*, and the word *тоска* – some parts of the meanings of the words *yearning, depression* and *anguish*).

The second source is different display rules: it seems that American English tends to avoid sending direct negative messages of the kind 'I did something bad' (prototypical setting for 'SHAME'), 'You made me feel bad' (prototypical setting for 'OFFENCE'), 'You feel bad' (prototypical setting for 'HURT FEELINGS'), 'You are in a bad situation' (prototypical setting for 'PITY') and replaces them with milder ones where possible, whereas there is no such constraint in Russian. It explains the relative higher frequency of the Russian terms for 'pity', 'shame', and 'offence', which has led to the inclusion of *жалость* 'pity', *неловко* 'I feel bad, I feel ashamed; lit.: uncomfortable', *обижать* 'to offend, to hurt', *обижаться* 'to feel hurt, offended' in the list of Russian ethno-specific key words [Levontina 2004], [Shmelev 2002], [Levontina & Zalizniak 2001]. On the whole, Russian discourse allows one to express **negative** feelings in a slightly exaggerated way, as compared to American English, which welcomes exaggerated expression of **positive** feelings (cf. much-discussed de-semanticized use of *happy* as compared to *счастливый*), but shuns the expression of negative feelings.

Thus, it is acceptable in the Russian cultural milieu to tell a person X that the experiencer feels sorry for X, or admit that (s)he

003-Apresjan_V:_Layout 1 18.05.2008 13:59 Page 20

feels offended by X, and it is polite to exaggerate one's feeling of shame before X for insignificant inconveniences caused by the experiencer. The English language, on the other hand, prefers to spare the feelings of both the experiencer and the object of emotion and slightly **diminish** them, often by using generalized terms like *to feel bad* in situations potentially embarrassing for either of the communicants; cf. *I feel bad for you, I feel for you* instead of *Мне тебя жалко* 'I feel pity for you'; *I feel bad to bother you* instead of *Мне неловко Вас беспокоить* 'I'm ashamed to bother you'*, This made me feel bad, I was sad* instead of *Мне было обидно* 'I was offended', *Take it easy* instead of *Не обижайся* 'Don't be offended'.

Below, are some excerpts from the comparison of 'FEAR' emotion clusters in Russian and English, from the total of 11 clusters analyzed. Many of the individual emotion terms mentioned above and below have received profound and thorough semantic descriptions in the works of [Iordanskaja 1971, 1984], [Iordanskaja et al. 1996], [Wierzbicka 1999], [Ju. D. Apresjan 2004], [Uryson 2004], [Levontina 2004], [Shmelev 2002], [Levontina & Zalizniak 2001] and other researchers.

'FEAR' cluster in Russian and English

Both languages have a concept of "general, neutral" 'FEAR', which is an emotional, rational and behavioral response to a potentially dangerous object, event, situation or action; at the prospect of coming into a closer contact with it, the experiencer wants to withdraw.

This type of 'FEAR' can be semantically explicated as follows

'a person X thinks that a person, event or action Y is dangerous; X wants to avoid Y; X feels bad': *I'm afraid of this man <of the exams>*; *I'm afraid to go into the woods*; *I'm scared to go down this slope*; *Я боюсь этого человека <экзаменов>*; *Я боюсь ходить в лес*; *Мне страшно ехать с горы.*

There is also a notion of short-lived "biological" 'FEAR', which does not involve any thinking prior to experiencing the emotion; it is the product of a sudden exposure to a frightening object or situation, such as a dog, a stranger, a sudden noise, etc. Semantically, it can be explicated as follows:

'X suddenly perceives an object Y; X-s body reacts to it as it reacts to danger - by lowered body temperature and heightened blood rate; X might do some involuntary actions as a result, for example, to run away or to freeze; X feels (s)he has experienced something bad'. This type of fear is expressed by several lexical items in English and Russian: *to get scared, to get frightened, to get a scare, to be shit-scared*, as well as the Russian *пугаться/испугаться/перепугаться*.

Both English and Russian have concepts for a very strong short-lasting 'FEAR'which can be explicated as 'X thinks that a very bad Y might happen or that a very bad and powerful Y might do something very bad to X; X feels that (s)he cannot do anything to prevent this; X feels very bad'.

In English, there is a special word to express this meaning, *terror* and its derivatives – *to terrify, terrifying, terrified*. This kind of 'FEAR' involves a very strong physiological and uncontrolled behavioral response, which is reflected in language, cf. *to go pale with terror, one's blood turns icy with terror, to freeze/to numb with terror; to flee in terror*. The Russian correlate of this word, *ужас*, as well as its derivatives, *ужасаться/ужаснуться* has a wider meaning: it can refer not only to the feeling of anticipating something very bad, but also to the feeling of being exposed to something very bad that has already happened. Thus, it is possible to use this word in both of the following contexts, where English uses two separate terms: *Он с ужасом смотрел на приближавшихся бандитов* 'He was looking at the approaching gangsters in terror' and *Он с ужасом смотрел на обезображенное тело* 'He was looking at the mutilated body in horror'. In the first usage, it is very much like *terror*; cf. *застыть <побелеть> от ужаса, быть парализованным ужасом, убежать в ужасе* 'to freeze with terror, to be paralyzed by terror, to go pale with terror, to flee in terror'.

In the second usage, it is different from *terror* and closer to *horror*, though it lacks the 'disgust' component of *horror.* The perfective verb form *ужаснуться*, derived from *ужас*, can only refer to the feeling that occurs after something bad has happened, not prior to it. The imperfective verb form *ужасаться* is also used to refer to something that has already happened, only it describes the verbal behavior of a person: *Он долго ужасался моему рассказу* 'lit: He for a long time was being terrified by my story' means 'For a while, he was expressing his horror at what I had told'.

Horror is an emotion which combines the elements of 'FEAR' and 'DISGUST'; it is a borderline emotion, a fact which is manifested even in its metaphorical conceptualization. While *horror* produces some typical *terror*-like reactions, it also involves some *disgust*-type reactions; cf. *horror iced <curdled> one's blood*; *to shrink in horror*, but also *to vomit in horror* (one cannot *vomit in terror* or in *ужас*). *Horror* is an example of removing barriers between emotion clusters, which is a very typical phenomenon for English. Thus, where in English we find *horror,* in Russian we may well find *ужас и отвращение* 'terror and disgust', or *ужас, смешанный с отвращением* 'terror mixed with disgust' to describe this particular brand of feeling. However, interestingly, though both 'strong fear' and 'disgust' components seem to be present in *horror*, their sum does not equal its meaning. In a very subtle description of *horror* [Solomon 2004] points out an important component of *horror* that seems to be absent in either of these two emotions: the component of breaking the norm, of shock from discovering, instead of something normal and familiar, something monstrous and ghastly.

Both languages have singled out a kind of religious 'FEAR' that is inspired by very powerful objects and forces like God or nature, though both for Russian and English it is a more marginal concept; thus, lexical items expressing it are less frequently used, and belong to a literary, rather than colloquial register.

This type of 'FEAR' is mixed with great respect and admiration, and the component of 'FEAR' in it is not enough to render the resulting feeling unpleasant. Y is not perceived as dangerous or potentially harmful; on the contrary, it is viewed as exceedingly good; however, Y's omnipotence as compared to the experiencer's relative weakness accounts for the 'FEAR'-component. Semantically, it can be explicated as follows:

'X thinks that Y is very powerful and very good; X feels something very good and very strong for Y; X feels that Y can do everything; X feels small and insignificant in the presence of Y; X cannot do anything in the presence of Y'.

In English, this type of *fear* is more pronounced: the word *awe* and its derivatives (*awesome, awed*) refer to this mixed type

003-Apresjan_V:_Layout 1 18.05.2008 13:59 Page 21

of feeling, whereas in order to express this combined meaning the Russian language would resort to using a phrase *благоговейный ужас* 'lit: awesome terror', or metaphorical expressions *трепет, трепетать* 'lit: quaking, to quake'. *Awe* does not contain 'FEAR' and 'ADMIRATION' in equal proportions, which is why *благоговейный <священный> ужас* is not an exact correlate; in *awe*, the measure of the good emotion, admiration, is stronger than that of the scary emotion; so, on the whole, it is perceived as a positive emotion rather than negative.

In contrast to this reverent, religious-like feeling, there is a totally rational type of 'FEAR' which involves little, if any, emotional components and implies a mostly rational appraisal of a certain object or situation as dangerous and, as a result, a controlled behavior in the form of consciously avoiding it:

'X thinks that Y is dangerous; X prefers to avoid Y'.

This type of 'FEAR' is expressed by *apprehensive* and its derivatives in English and *опасаться* and its derivatives in Russian: *Apprehensive about the side effects of anti-depressants, he opted out of pharmacotherapy*; *Опасаясь побочных эффектов антидепрессантов, он отказался от медикаментозного лечения*.

An absolute opposite to the rational 'FEAR' are *panic* and *freaking out*, as well as the Russian *паника*. This type of 'FEAR' implies complete loss of rational control over emotions and, in the case of *panic* and *паника*, uncontrolled behavioral reactions; unlike all other types of 'FEAR', this one can characterize the psychotic behavior of large groups of people, even crowds; cf. the psychological term *crowd panic.*

Since 'FEAR' involves behavioral responses, its appraisal is partly triggered by social and ethical norms. Both languages contain a concept of "bad, unethical" 'FEAR' or, rather, unethical behavior in the situation when a person experiences fear. The situations themselves might differ with time, place and culture, but there are always some which require bravery, and the failure to live up to the required expectations results in negative ethical evaluation of the person and his (her) behavior. This type of 'FEAR' can be explicated as

'X feels that Y is dangerous; X wants to avoid Y; X tries to avoid Y; the speaker thinks avoiding Y is bad'.

This type of behavior and type of personality associated with it is expressed by the English *to get cold feet*, *coward, cowardly, chicken* and the Russian *(с)трусить* and *(с)дрейфить* , where the imperfective form implies reluctance to do something and the perfective form – a complete withdrawal from the situation.

Characteristically, both languages employ the same metaphorical means to describe this cowardly behavior in a derogatory way, which are based on one of the rarer symptoms of fear – involuntary defecation; cf. *to poop out, to crap out* or the Russian expression *наложить в штаны* which all mean 'X didn't do something because of fear; the speaker thinks this is very bad' [cf. Dobrovol'skij 1996 on this expression in Russian and German]. Another way of carrying negative evaluation metaphorically which is often used in the field of emotions is likening the experiencer's behavior to that of an animal; cf. the expression with a similar meaning *to have one's tail between one's legs* or its Russian equivalent *поджать хвост*.

References

1. Apresjan & Apresjan 1993 – Ju.D.Apresjan, V.Ju.Apresjan. Metafora v semantičeskom predstavlenii emocij // Voprosy jazykoznanija, n.3, pp. 27-35.

2. Apresjan 2000 – Apresjan Juri, Systematic Lexicography // Translated by Kevin Windle. Oxford: Oxford University Press, 2000.

3. Apresjan 2004 – Ju.D.Apresjan. Sinonimičeskie riady GORDIT'SIA, SERDIT'SIA, RADOVAT'SIA // Novyj Objasnitel'nyj Slovar sinonimov russkogo jazyka. Izdanie vtoroe, ispravlennoe i dopolnennoe. Pod obščim rukovodstvom akademika Ju.D.Apresjana. Moscow 2004.

4. Boym 1994 – Svetlana Boym. Common Places: Mythologies of Everyday Life in Russia. Cambridge. Harvard University Press, 1994.

5. Bulygina, Shmelev 1997 – T.V.Bulygina, A.D. Shmelev. Jazykovaja konceptualizacija mira na materiale russkoj grammatiki. Moscow, 1997.

6. Calder et al. 2001 – Andrew J. Calder, Andrew J. Lawrence, Andrew W. Young. Neuropsychology of Fear and Loathing // Neuroscience, May 2001, vol. 2. pp. 352-363.

7. Clanton & Smith 1977 - Gordon Clanton, Lynn G. Smith, ed. Jealousy. Prentice Hall, 1977.

8. Dobrovol'skij 1996 – D.O.Dobrovol'skij. Obraznaja sostavljajuščaja v semantike idiom // Voprosy jazykoznanija1996, N 1.

9. Ekman 1984 – Paul Ekman. Expression and the Nature of Emotion. In: Approaches To Emotion. Ed.: Lawrence Erlbaum Associates, Publishers, 1984.

10. Ekman 1999 – Paul Ekman. Basic Emotions // In T. Dalgleish and M. Power (Eds.). Handbook of Cognition and Emotion. Sussex, U.K.: John Wiley & Sons, Ltd., 1999. Chapter 3.

11. Emanatian 1995 – Michele Emanatian. Metaphor and the Expression of Emotion: The Value of Cross-Cultural Perspectives // Metaphor and Symbolic Activity. 10(3), pp. 163-182. 1995.

12. Gallese 2001 – Vittorio Gallese. The "shared manifold" hypothesis: from mirror neurons to empathy // Journal of Consciousness Studies: 8, N 5-7, 2001. pp. 33-50.

13. Iordanskaja 1971 – L.N.Iordanskaja. Leksikografičeskoe opisanie russkix vyraženij, oboznačajuščix fizičeskie simptomy čuvstv // Mašinnyj perevod i prikladnaja lingvistika, volume 16, 1972.

14. Iordanskaja 1984 – L.N.Iordanskaja. Slovarnye statji bojat'sia, vostorg, vosxiščat', gnev, strax // Tolkovo-kombinatornyj slovar' sovremennogo russkogo jazyka. Vienna 1984.

15. Iordanskaja et al. 1996 - Lidija Iordanskaja, Slava Paperno, Richard L. Leed. A Russian-English Collocational Dictionary of

003-Apresjan_V:_Layout 1 18.05.2008 13:59 Page 22

the Human Body. Columbus/Ohio, 1996.

16. Jabbi, Swart & Keysers 2007 – Mbemba Jabbi, Marte Swart, Christian Keysers. Empathy for Positive and Negative Emotions in the Gustatory Cortex // NeuroImage 34, 2007. pp. 1744-1753.

17. Kövecses 1990 – Zoltan Kövecses. Emotion Concepts. Frankfurt-am-Main. Springer Verlag, 1990.

18. Kövecses 2000 - Zoltan Kövecses. Metaphor and Emotion. Cambridge University Press, 2000.

19. Lakoff & Johnson, 1980 – George Lakoff, Mark Johnson. Metaphors we live by. Chicago & London, The University of Chicago Press, 1980.

20. LeDoux 1996 – Joseph LeDoux. The Emotional Brain. Simon and Schuster, 1996.

21. Levontina 2004 – I. B. Levontina. Sinonimičeskij riad ŽALOST' // Novyj Objasnitel'nyj Slovar sinonimov russkogo jazyka. Izdanie vtoroe, ispravlennoe i dopolnennoe. Pod obščim rukovodstvom akademika Ju.D.Apresjana. Moscow 2004.

22. Levontina & Zalizniak 2001 – Irina Levontina, Anna Zalizniak. Human Emotions Viewed through Russian Language // Emotions in Cross-Linguistic Perspective. Ed. by Jean Harkins, Anna Wierzbicka. Mouton de Gruyter, 2001.

23. MacDaniel 1996 – Tim MacDaniel. The Agony of the Russian Idea. Princeton University Press, 1996.

24. McDonald & Leary 2005 – Geoff McDonald, Mark R. Leary. Why Does Social Exclusion Hurt? The Relationship Between Social and Physical Pain // Psychological Bulletin 2005, Vol. 131, No. 2, pp. 202–223.

25. Mel'čuk & Wanner 1996 – Igor Mel'čuk, Leo Wanner. Lexical Functions and Lexical Inheritance for Emotion Lexemes in German // Lexical Functions in Lexicography and Natural Language Processing, ed. by Leo Wanner. John Benjamins, Amsterdam, 1996.

26. Oatley & Jenkins 1996 – Keith Oatley, Jennifer M. Jenkins. Understanding Emotions. Blackwell Publishers, 1996.

27. Rancour-Laferriere 1995 – Daniel Rancour-Laferriere. The slave soul of Russia: moral masochism and the cult of suffering. New York New York University Press,1995.

28. Rougemont 1956 – Denis de Rougemont. Love in the Western World. Princeton University Press, 1956.

29. Scherer 1988 – Klaus Scherer, ed. Facets of Emotions: Recent Research. Lawrence Erlbaum Associates, 1988.

30. Scherer, Wallbott & Summerfield 1987 – Klaus R. Scherer, Harald G. Wallbott & Angela B. Summerfield. Experiencing Emotion: A Cross-Cultural Study. Cambridge University Press, 1987.

31. Shmelev 2002 – A. D. Shmelev. Russkij jazyk i vnejazykovaja dejstvitel'nost'. Moscow, 2002.

32. Shweder 2004 – Richard A. Shweder. Deconstructing the Emotions for the Sake of Comparative Research // Feelings and Emotions. Ed. by Anthony Manstead, Nico Frijda, Agneta Fischer. Cambridge University Press, 2004.

33. Singer et al. 2004 – Tania Singer, Ben Seymour, John O-Doherty, Holger Kaube, Raymond Dolan, Chris Frith. Empathy for Pain Involves the Affective but not Sensory Components of Pain // Science, 2004, vol. 303, issue 5661, pp. 1157-1162.

34. Solomon 2004 – Robert Solomon. In Defense of Sentimentality. Oxford University Press, 2004.

35. Uryson 2004 – E.V. Uryson. Sinonimičeskij riad TOSKA // Novyj Objasnitel'nyj Slovar sinonimov russkogo jazyka. Izdanie vtoroe, ispravlennoe i dopolnennoe. Pod obščim rukovodstvom akademika Ju.D.Apresjana. Moscow 2004.

36. Wierzbicka 1990 – Anna Wierzbicka. Dusa `soul', toska `yearning', sud'ba `fate': three key concepts in Russian language and Russian culture. In: Metody formalne v opisie językow slowiańskix, ed. Zygmunt Saloni. Dzial Wydawnitctw Filii UW w Bialymstoku, 1990.

37. Wierzbicka 1991 – Anna Wierzbicka. Cross-cultural Pragmatics: the Semantics of Social Interaction. Berlin: Mouton de Gruyter, 1991.

38. Wierzbicka Anna, 1992 – Anna Wierzbicka. Semantics, Culture and Cognition. Universal Human Concepts in Culture-Specific Configurations. Oxford University Press, 1992. New York, Oxford.

39. Wierzbicka Anna, 1997 – Anna Wierzbicka. Understanding Cultures through their Key Words. Oxford University Press, 1997.

40. Wierzbicka 1999 – Anna Wierzbicka. Emotions across Languages and Cultures. Cambridge University Press, 1999.

41. Wierzbicka 2001 - Anna Wierzbicka. Introduction // Emotions in Cross-Linguistic Perspective. Ed. by Jean Harkins, Anna Wierzbicka. Mouton de Gruyter, 2001.

004-Apresjan_J:_Layout 1 18.05.2008 14:19 Page 23

О ПРОЕКТЕ АКТИВНОГО СЛОВАРЯ (АС) РУССКОГО ЯЗЫКА[1]
ON A PROJECT OF A PRODUCTION DICTIONARY OF RUSSIAN

Апресян Ю.Д. *(apr@iitp.ru)*
Институт русского языка им. В.В.Виноградова РАН

Доклад посвящен проекту активного словаря русского языка, работа над которым ведется с 2006 г. в ИРЯ РАН. Словарь должен аккумулировать достижения европейской активной лексикографии и результаты современных исследований в области семантики (теория толкований и семантические правила), синтаксиса (валентные и невалентные свойства лексем), сочетаемости (теория лексических функций) и лексикализованной просодии.

1. Основные формулы пассивного и активного словарей

Основная формула пассивного словаря – много слов, минимальная информация о каждом слове, достаточная для его понимания в произвольном контексте. О слове *мнение*, например, достаточно сообщить, что это – оценочное суждение о чем-л., основанное на опыте, вкусах или умозаключениях субъекта. В контексте слово будет понятно.

Основная формула активного словаря – существенно меньше слов, но по возможности полная информация о каждом слове, необходимая для его правильного употребления в собственной речи говорящих. При таком подходе в фокусе внимания лексикографа чаще всего оказывается не вполне свободная сочетаемость слова. Она представляет наибольшие трудности и при усвоении родного языка, и при изучении иностранного языка. Для того же слова *мнение* в активном словаре следовало бы привести: 1) сочетания с прилагательными, например, *твердое <сложившееся> мнение, высокое <лестное, положительное> мнение, невысокое <низкое, нелестное, отрицательное> мнение, субъективное <предвзятое, спорное> мнение, объективное <непредвзятое> мнение, общественное <личное> мнение, общее <единое> мнение*; 2) сочетания с существительными, например *столкновение <борьба> мнений, разноречивость <пестрота> мнений*, *смена мнений*, *центр изучения общественного мнения*; 3) сочетания с глаголами, при которых *мнение* играет роль дополнения, например, *иметь мнение, быть какого-то мнения, держаться мнения, приходить к мнению, выражать свое мнение, оставаться при своем мнении, отказываться от мнения, отвергать <разделять> чье-л. мнение, менять мнения*; 4) сочетания с глаголами, при которых *мнение* играет роль подлежащего, например, *Есть мнение (что Р)*, *Мнение складывается <создается, укореняется>, Мнение изменяется, Мнения сталкиваются <расходятся, сходятся, совпадают>*.[2]

Современная европейская, особенно английская лексикография располагает большим семейством активных словарей; из последних по времени назову Oxford 2003, Longman 2003, Macmillan 2002, содержащих до 40-50 тыс. слов. В нашей лексикографии такие словари представлены только экспериментальными изданиями типа Мельчук и Жолковский 1984 (около 300 словарных статей) и Денисов, Морковкин 1978 (около 2500 слов). Пришло время заполнить эту лакуну.

Я расскажу о проекте АС русского языка, работа над которым началась в ИРЯ РАН в 2006 г. Приступая к этому проекту, мы планируем не просто воспроизвести на русском материале тип активного словаря, сложившийся в европейской традиции, а радикально его модернизировать с опорой на современные лингвистические технологии и теории.

Ввиду ограниченности места я вынужден буду иллюстрировать свои тезисы выборочно – на единичных, но представительных примерах. Лейтмотивом всех примеров будет идея системного описания лексики, или идея (высоковероятных) предсказаний, которые можно делать на основании отдельных свойств лексем или их при-

[1] Эта работа была поддержана грантом РГНФ № 06-04-00289а, грантом РФФИ № 05-06-80361, грантом Программы фундаментальных исследований ОИФН РАН «Русская культура в мировой истории» и грантом Президента РФ для поддержки научных исследований, проводимых ведущими научными школами РФ, № НШ-5611.2006.6. Она построена на основе доклада, прочитанного 22.10.07 в Москве на Международной конференции «Русский язык в странах СНГ и Балтии», но материал ее существенно обновлен.

[2] Бо'льшая часть этих сочетаний содержится в словарной статье МНЕНИЕ (авторы – Ю.Д. Апресян, А.К. Жолковский, И.А. Мельчук); см. Мельчук и Жолковский 1984: 424-432.

004-Apresjan_J:_Layout 1 18.05.2008 14:19 Page 24

надлежности к тем или иным семантическим классам и подклассам. Приводимый материал обладает еще одним общим свойством – все это факты, которые в существующих словарях либо вовсе не отражаются, либо представлены неполно.

2. Современные технологии

Здесь я имею в виду опору на корпусы и лингвистический эксперимент.

2.1. Корпусы текстов

Полезность корпусных данных для лингвистических исследований сейчас уже не требует доказательств. Я ограничусь одним простым примером, показывающим, какие новые знания, полезные и теоретически, и практически, можно извлечь из корпусов текстов.

В русской лексикографии, начиная с Академического словаря 1847, слова *ноль* и *нуль* трактуются как совершенно равноправные варианты, причем до СУш 1934-1940 основным вариантом считался *нуль*, а в этом словаре и после него – *ноль*.

Более тонко они описаны в Орфоэпическом словаре 1989, где предлагается следующее распределение по формам: в форме ИМ *ноль* нейтрален, а *нуль* считается его допустимым, но устаревшим вариантом. Обычно говорят *один – ноль в нашу пользу*, а не *один – нуль в нашу пользу*, *У нас – ноль градусов*, а не *нуль градусов* и т.п. Напротив, во всех косвенных падежах нейтрален *нуль*, а *ноль* считается его допустимым вариантом. Обычно говорят *равный нулю*, а не *равный нолю*, *начинать с нуля*, а не *начинать с ноля*.

В Зализняк 2003 замечены еще три важные детали распределения: а) в математическом значении обычен вариант *нуль*; б) в прочих значениях в формах ИМ ЕД и ВИН ЕД употребляется преимущественно вариант *ноль*; в) все остальные формы образуются преимущественно от варианта *нуль*.

Эти наблюдения тем более поразительны, что сделаны в до-корпусную эпоху, исключительно на основании выдающейся языковой и лингвистической интуиции автора словаря[3]. Однако обращение к текстам позволяет внести и в эту картину еще два уточнения. С одной стороны, даже в математических текстах в названиях дробей в формах ИМ и ВИН предпочитается вариант *ноль*: *ноль целых две десятых*, *на ноль целых две десятых* лучше, чем *нуль целых две десятых*, *на нуль целых две десятых*. Здесь форма (винительный падеж) оказывается сильнее, чем значение. С другой стороны, вариант *нуль* является показателем не только математического, но и любого другого специального дискурса, т.е. тяготеет к закреплению в роли термина. Соответственно он предпочитается также в тех случаях, в которых имеет место метафорическое переосмысление первоначально терминологических сочетаний или образов; ср. *абсолютный нуль*, *В итоге полный <абсолютный> нуль*; *Вместо плюса нуль – у нас минус нуль* (Е. Замятин, Мы); *Типичная нуль-транспортировка* (А. и Б. Стругацкие, Попытка к бегству); *При некоторой температуре Тс, называемой точкой Кюри, <...> намагниченность точно обращается в нуль* (Уппсальский корпус).

Таким образом, в современном русском языке сложились новые правила употребления вариантов *ноль* и *нуль*, пока еще не описанные ни в каких нормативных источниках. Они имеют прямое отношение к формированию навыков правильной русской речи и поэтому в каком-то виде должны быть включены в АС русского языка.

2.2. Лингвистический эксперимент

Хочу сказать два слова и в защиту лингвистического эксперимента. Он был в большой чести у теоретиков в эпоху господства структурализма, часто в ущерб опоре на хорошие тексты, но сейчас, когда нам стали доступны громадные языковые материалы корпусов, он стал вызывать скепсис у поклонников корпусной лингвистики. Между тем в текстах встречаются употребления, явным образом противоречащие сложившейся языковой норме. Чтобы отличить такие употребления от реально скрытых в языке возможностей, часто бывает необходимо прибегнуть к лингвистическому эксперименту и сопровождающему его объяснительному инструментарию.

В качестве примера я сначала рассмотрю лексическую единицу *в бытность*, которая стала предметом спе-

[3] Приведем весьма яркие данные поисковой системы GOOGLE по формам множественного числа, подтверждающие наблюдения А.А. Зализняка: ИМ и ВИН: *ноли* – 56 000 употреблений, *нули* – 943 000 употреблений; РОД: *нолей* – 27 700 употреблений, *нулей* – 357 000 употреблений; ДАТ: *нолям* – 643 употребления, *нулям* – 122 000 употреблений; ТВОР: *нолями* – 23 500 употреблений, *нулями* – 270 000 употреблений; ПР: *нолях* – 1880 употреблений, *нулях* – 48 000 употреблений. Полезно обратить внимание на то, что, несмотря на огромное числовое превосходство *нуля* над *нолем*, основным вариантом в современном русском языке остается *ноль*: представляющая форма существительного, т.е. ИМ ЕД, оказывается более сильным фактором, чем употребительность.

004-Apresjan_J:_Layout 1 18.05.2008 14:19 Page 25

циального исследования в работе В.С. Храковского, посвященной таксису[4]. Среди сделанных автором наблюдений было следующее: в корпусе, которым он пользовался, эта единица встретилась 107 раз в контексте местоимений первого лица (*в мою <нашу> бытность*), 83 раза в контексте местоимений третьего лица (*в его <ее, их> бытность*) и только 6 раз в контексте местоимений второго лица (*в твою <вашу> бытность*). В.С. Храковский эти предпочтения в способах заполнения первой валентности единицы *в бытность* зафиксировал, но не дал им объяснения. Между тем объяснение возможно. Дело в том, что эта единица стилистически маркирована как нарративная и книжная. Она звучит вполне естественно в рассказе о событиях своей собственной жизни или жизни третьих лиц. Однако в контекстах типа *В твою <в вашу> бытность председателем правления порядка было больше* возникает небольшой, но заметный конфликт между нарративностью, повествовательностью самой единицы и диалогичностью высказывания. Хотя, таким образом, в текстах есть употребления типа *В твою <вашу> бытность*[5], их следует квалифицировать, с учетом сказанного, как слегка отклоняющиеся от нормы.

Итак, наличие каких-то фактов в текстах еще не является свидетельством того, что они являются и фактами языка и, тем самым, подлежат включению в нормативные грамматики и словари (а активный словарь не может не быть нормативным). Необходим эксперимент, использующий непосредственную языковую интуицию говорящих по поводу того, что правильно, а что неправильно в языке, результаты которого должны быть объяснены с точки зрения современной лингвистической теории. Только после этого принимается окончательное решение о статусе данного явления в языке и о том, надо ли его включать в словарь, и если да, то с какой пометой.

Рассмотрим еще одну ситуацию, характерную для работы лексикографа и тоже требующую обращения к эксперименту и технике лингвистических объяснений. Возьмем группу глаголов а) *бахвалиться*, *хвалиться*, *хвастаться* и б) *рисоваться*, *форсить*, *щеголять* в близких значениях ‘сообщать кому-л. о своих достоинствах, сильно их преувеличивая’, ‘вести себя так, чтобы произвести на кого-то впечатление своими достоинствами’ (уточненный пример из работы Апресян 1992). У всех этих глаголов, по данным современных словарей русского языка, имеется свойство управлять существительным в форме ТВОР: *бахвалиться <хвалиться*, *хвастаться> своими успехами*, *рисоваться <форсить*, *щеголять> своей силой.* В ряде словарей (например, в МАСе) отмечается, кроме того, способность глаголов *хвалиться* и *хвастаться* управлять предложно-именной группой *перед кем-л.*: *хвалиться <хвастаться> перед кем-л. своими успехами*. Однако ни МАС, ни БАС, ни другие авторитетные толковые словари русского языка не содержат аналогичных указаний относительно глаголов *бахвалиться*, *рисоваться*, *форсить* и *щеголять.* Возникает вопрос, является ли это принципиальным фактом, отражающим закономерность языка, или случайным обстоятельством, возникшим вследствие отсутствия таких примеров в словарных картотеках.

Обращение к современным корпусам дает приблизительно одинаковую картину для всех этих глаголов. По данным текстов, все они способны к двойному управлению типа *бахвалиться <хвалиться*, *хвастаться> чем-л. перед кем-л.*, *рисоваться <форсить*, *щеголять> чем-л. перед кем-л.*

Чтобы понять, в какой мере можно доверять конфликтующим данным авторитетных словарей и корпусов, следует обратиться к языковой интуиции говорящих (т.е. в конечном счете к эксперименту) и попытаться найти в языке обоснование для нее.

Как будто нет оснований сомневаться в том, что в данном случае, в отличие от рассмотренного выше, правы тексты: обе группы словосочетаний с точки зрения носителя языка вполне правильны. Можно предложить и семантическое объяснение этой интуиции.

Из приведенных выше толкований следует, что глаголы первой группы обозначают разновидности сообщения о своих достоинствах, а глаголы второй группы – разновидности демонстрации своих достоинств. Но сообщение и демонстрация предполагают Аудиторию (слушателей или зрителей), для которых они предназначены. Именно эту роль реализует форма *перед кем-л.*

Толкования дают основание для постулирования еще некоторых управляющих свойств глаголов первой группы. *Бахвалиться*, *хвалиться* и *хвастаться* суть речевые акты, а речевые акты предполагают, во-первых, Содержание сообщения, которое может выражаться придаточным предложением с союзом *что*, и, во-вторых, Адресата, который обычно выражается именной группой в форме ДАТ. Такие группы легко подставляется в словосочетания со всеми тремя глаголами. Ср. *Он бахвалился <хвалился*, *хвастался> мне, что может одним ударом свалить быка.* Обоими этими свойствами они отличаются от глаголов поведения *рисоваться*, *форсить* и *щеголять.* Надо сказать, что эти два свойства, легко устанавливаемые в эксперименте, в толковых словарях современного русского языка либо вовсе не отмечаются, либо отмечаются крайне непоследовательно.

[4] Мои соображения основаны на материалах доклада «Нестандартный таксис: анализ конструкций с грамматической единицей *в бытность*», прочитанного В.С. Храковским 19 октября 2007 г. на Семинаре по теоретической семантике в ИППИ РАН.
[5] В GOOGLE’е их обнаружилось около 600 000 – против 6 569 000 с местоимениями первого и третьего лица (учитывались пре- и постпозиция местоимения относительно существительного).

004-Apresjan_J:_Layout 1 18.05.2008 14:19 Page 26

3. Теоретические результаты

Что касается теоретических результатов, то для АС особенно важны следующие четыре области лингвистической теории: а) семантика – теория толкований и правила взаимодействия значений; б) синтаксис – трехуровневая теория управления и аппарат для описания невалентных синтаксических свойств лексем; в) сочетаемость – обновленная теория лексических функций; г) просодия, где лексикографический интерес представляют факты лексикализации фразовых акцентов и их связь с коммуникативной организацией высказывания.

3.1. Семантика в АС

3.1.1. Толкования

К толкованиям, помимо общенаучных требований системности, полноты, неизбыточности и нетавтологичности, в АС предъявляются еще два требования. Во-первых, толкования должны быть доступны для непрофессионала. Во-вторых, они должны обладать объяснительной силой, например, объяснять допустимость одних словосочетаний и недопустимость других. Поясню эти требования на примере основных значений глаголов *войти* (*в комнату*) и *выйти* (*из комнаты*).

Толкования этих лексем в традиционных словарях выглядят следующим образом: *войти* = ‘идя, двигаясь, проникнуть куда-л., в пределы чего-л.’, *выйти* = ‘уйти откуда-л., оставить пределы чего-л.’, причем *уйти* в свою очередь толкуется как ‘покинуть какое-л. место, чье-л. общество; удалиться, отправиться куда-л.’.

Эти толкования, бесспорно, доступны для непрофессионала, но никакому другому из перечисленных выше естественных требований они не удовлетворяют. В частности, они не объясняют, почему по-русски можно сказать *войти в дом с улицы, войти в комнату из холла*, а **войти из дома на улицу*, **войти из комнаты в холл* – нельзя; почему можно сказать *выйти из дома на улицу*, *выйти из комнаты в холл*, а **выйти в дом с улицы*, **выйти в комнату из холла* нельзя.

Дело в том, что глаголы *войти* и *выйти*, как и другие глаголы перемещения с приставками *в-* и *вы-* (а это – большой системный класс) накладывают определенные ограничения на пространства, где кто-то находился до начала перемещения и после него. Глаголы с приставкой *в-* обозначают перемещение в более замкнутое, а глаголы с приставкой *вы-* – перемещение в более открытое пространство. Уточненные таким образом толкования вполне объясняют, почему фразы *войти в дом с улицы* и *выйти из дома на улицу* воспринимаются носителями языка как правильные, а фразы **войти из дома на улицу* или **выйти в дом с улицы* – как неправильные.

3.1.2. Семантические правила

Перейду к правилам. Интересные семантические особенности имеют краткие формы некоторых параметрических прилагательных, прежде всего прилагательных линейного размера. Это слова *высокий – низкий*, *глубокий – мелкий*, *длинный – короткий*, *широкий – узкий*, а также о прилагательные «общего размера» *большой* и *маленький*. Назову две их особенности; обе были коротко описаны в [Апресян 1974: 93, 214].

Во-первых, прилагательные со значением большого полюса в контексте вопросительных слов *как* и *насколько* утрачивают присущий им в других конструкциях семантический компонент ‘больше нормы’ и становятся обозначениями всей шкалы соответствующего линейного измерения, т. е. семантическими эквивалентами слов *высота*, *длина*, *ширина*, *глубина* и т. п. Ср. вопросы типа *Насколько высок забор в этом месте? ≈ Какова высота забора в этом месте?*, *Насколько <как> длинны каналы Марса? ≈ Какова длина каналов Марса?*, *Насколько широк <глубок> ручей в нижнем течении? ≈ Какова ширина <глубина> ручья в нижнем течении?* Кстати, точно так же ведут себя и некоторые наречия со значением большого полюса, например, наречие *часто*: *Насколько часто вы чувствуете себя виноватым*, *Как часто вы думаете о сексе* (вопросы из психологических тестов).

Во-вторых, большинство параметрических прилагательных со значением линейного размера, включая и прилагательные малого полюса, равно как и оба прилагательных «общего размера» в краткой форме претерпевают семантический сдвиг вида ‘больше нормы Р’ ⇒ ‘слишком большой по Р’. Если в первом случае словарное толкование лексемы теряет некий компонент, то во втором случае оно приобретает новый компонент. Ср. *Забор высок <низок>*, *Кровать широка* (*сюда не встанет*), *Протока узка <мелка>* (*на катере не пройдешь*), *Лестница низка* (*до крыши не достанет*), *Леска тонка* (*на крупную рыбу не годится*), *Юбка коротка <длинна>*, *Ботинки малы <велики>, Мешки тяжелы* (*один человек не справится*) и т.п. Прилагательные других семантических классов сохраняют свои обычные значения и в краткой форме; ср. *Он красив <умен*, *добр, талантлив>*, *Замысел его глубок, но трудно осуществим, Он прочен, мой азийский дом.*

Этот сдвиг характерен для кратких форм при обычном порядке слов и вне контекста интенсификаторов. В положении перед подлежащим или при наличии интенсификатора краткие формы сохраняют свое словарное значение. Так обстоит дело в предложениях типа *Широка страна моя родная*, *Велика Россия, а отступать некуда*,

004-Apresjan_J:_Layout 1 18.05.2008 14:19 Page 27

Мал золотник, да дорог, Иван очень высок.

На основе этих и других подобных фактов было введено понятие употребления, которое стало основным инструментом описания семантических модификаций прототипического значения лексемы. Особым употреблением данного лексического значения называется такой сдвиг в его словарном толковании, который в строго верифицируемых контекстуальных условиях происходит всегда и поэтому может быть получен с помощью стопроцентно справедливого семантического правила.

3.2. Синтаксис в АС

3.2.1. Трехуровневая теория управления

Главным инструментом описания управляющих свойств предикатных лексем в АС является модель управления (МУ) – лексикографический конструкт, введенный в теории «Смысл ⇔ Текст» И.А. Мельчука. С помощью МУ описываются попарные соответствия между объектами следующих трех уровней: семантического, глубинно-синтаксического и поверхностно-синтаксического.

Отправной точкой для построения МУ любого предиката служит аналитическое толкование, позволяющее установить число его семантических актантов. Оно равно числу используемых в толковании переменных A1, A2, …An, а это число, в свою очередь, определяется составом обязательных участников ситуации, описываемой этим предикатом. Например, в ситуации, описываемой глаголом *прибивать / прибить* в его основном значении, участвуют пять объектов: 1) тот, кто прибивает (A1, Агенс), 2) то, что прибивается (A2, Пациенс), 3) то, к чему прибивается (A3, Пациенс / Место), 4) то, с помощью чего прибивается (A4, инструмент), 5) то, чем прибивается (A5, средство)[6]. Толкование: ‘Человек A1 прикрепил предмет A2 к предмету A3 ударами молотка A4 по гвоздю или гвоздям A5’ [в функции гвоздей и молотка могут использоваться и другие подобные им предметы].

Управление усматривается только в том случае, когда семантические актанты предиката реализуются словами или группами слов, которые подчиняются ему и синтаксически. В соответствии с этим нетривиальная часть модели управления (МУ) глагола *прибивать* / *прибить* будет выглядеть следующим образом:

A2, ВИН: *прибивать дощечку <планку, каблук, подметку, половицу>*;

A3.1, *к* ДАТ: *прибивать к стене <к полу, к доске объявлений, к двери>*;

A3.2, *на* ВИН: *прибить на дверь* (*табличку*) [A3 обычно вертикально ориентирован];[7]

A4, ТВОР: *прибивать топориком <сапожным молотком>*;

A5, ТВОР: *прибивать оцинкованными гвоздями*.

Однако только аналитическим толкованием дело не ограничивается. В Апресян 2006 я попытался показать, что в сложных случаях бывает необходимо выйти за пределы этого традиционного инструментария модели «Смысл ⇔ Текст» и использовать более общие соображения, связанные с устройством семантической системы языка в целом, в особенности с семантическими классами и подклассами фундаментальной классификации предикатов.

Один из таких сложных вопросов – вопрос об упорядочении актантов: почему актанту Инструмент в МУ *прибивать* приписан более высокий ранг, чем актанту Средство, а актанту Пациенс / Место – более высокий ранг, чем актанту Инструмент?

Самый общий ответ состоит в том, что в системе языка ранг актанта Ai, имеющего семантическую роль R, определяется числом предикатов, у которых он есть: чем больше число таких предикатов, тем выше ранг Ai. Рассмотрю с этой точки зрения конкуренцию «Инструмент – Средство».

Глагол *прибивать* / *прибить* входит в класс пятиместных предикатов с приставкой *при-*, обозначающих действия, для выполнения которых нужны Инструменты и Средства: *прибить*, *привинтить*, *приклепать*, *приколотить*, *приметать, припаять, пристегать* (*подкладку*), *пристрочить, притачать* (*голенище*), *пришить.* По объему с этим классом сравним семантически близкий класс четырехместных предикатов с приставкой *при-*, у которых есть актант со значением Средства, но нет актанта со значением Инструмента: *привязать*, *приклеить*, *приковать* (*себя наручниками к перилам*), *приколоть*, *прилепить*, *прикрутить*, *приметать*, *примотать*, *при-*

---

[6] Пояснения о семантических ролях. Сдвоенная роль Пациенс / Место приписывается такому участнику ситуации, который является местом, изменяющим в ходе действия или процесса свое состояние или свойства. У глагола *привить (ветку к яблоне)* есть актант с такой ролью [яблоня в результате прививки изменяет свои свойства], а у глаголов п*ригнуть (ветку к земле), придвинуть (стул к шкафу), прижать (руку к груди), приникнуть (к земле), прислонить <приставить> (лестницу к дереву)* так же оформленный актант выполняет семантическую роль Конечной точки [земля, шкаф и т.п. не меняют ни состояния, ни свойств]. Средство – роль такого вспомогательного объекта, который расходуется в процессе выполнения действия (гвозди, нитки, зубная паста, пули, стрелы и т.п.), а Инструмент – роль такого вспомогательного объекта, который не расходуется (молоток, иголка, зубная щетка, пистолет, лук и т.п.).

[7] **Прибить табличку на стол* звучит существенно хуже, чем *прибить табличку на дверь.*

004-Apresjan_J:_Layout 1 18.05.2008 14:19 Page 28

стегнуть, *прицепить*, *пришпилить*. Сравнимого класса глаголов с приставкой *при-*, у которых был бы актант Инструмент без актанта Средство, мы не знаем. Таким образом, в пределах этого класса Средство как будто получает преимущество перед Инструментом.

Картина радикально меняется, как только мы выходим за пределы класса глаголов с приставкой *при-*: класс четырехместных предикатов с актантом Инструмент оказывается в четыре-пять раз больше сравнимого по объему класса предикатов с актантом Средство.

Возьмем, например, корневые глаголы, обозначающие конструктивные физические действия, для выполнения которых необходимы инструменты с острым краем или концом. Вот неполный список глаголов этого класса: *брить*, *бурить*, *долбить*, *колоть*, *копать*, *косить*, *пилить*, *резать*, *рубить*, *рыть*, *сверлить*, *скоблить*, *стричь*, *строгать*, *тесать*, *точить*. К ним по разным признакам близки еще два глагола – *бить* и *колотить*.

От большинства этих глаголов с помощью приставок *в-*, *вы-*, *за-*, *от- <ото->*, *про-*, *с- <со->* и, может быть, еще некоторых образуется от одного до шести производных четырехместных предикатов со следующей актантной структурой: Агенс, Пациенс, Пациенс / Место, Инструмент. Ср. *вбить* (*что, во что, чем*), *выбить* (*что, из чего, чем*), *забить* (*что, во что, чем*), *отбить* (*что, от чего, чем*), *пробить* (*что, в чем, чем*), *сбить* (*что, с чего, чем*); *выбрить* (*что, на чем, чем*), *пробрить* (*что, в чем, чем*), *сбрить* (*что, с чего, чем*); *вкопать* (*что, во что, чем*), *выкопать* (*что, из чего, чем*), *закопать* (*что, во что, чем*), *откопать* (*что, из чего, чем*), *прокопать* (*что, в чем, чем*); *выпилить* (*что, из чего, чем*), *отпилить* (*что, от чего, чем*), *пропилить* (*что, в чем, чем*), *спилить* (*что, с чего, чем*); *врезать* (*что, во что, чем*), *вырезать* (*что, из чего, чем*), *вырезать* (*что, на чем, чем*), *отрезать* (*что, от чего, чем*), *прорезать* (*что, в чем, чем*), *срезать* (*что, с чего, чем*); *вырубить* (*что, из чего, чем*), *вырубить* (*что, на чем, чем*), *отрубить* (*что, от чего, чем*), *прорубить* (*что, в чем, чем*), *срубить* (*что, с чего, чем*); *врыть* (*что, во что, чем*), *вырыть* (*что, из чего, чем*), *зарыть* (*что, во что, чем*), *отрыть* (*что, из чего, чем*), *прорыть* (*что, в чем, чем*); *высверлить* (*что, в чем, чем*), *просверлить* (*что, в чем, чем*), *рассверлить* (*что, в чем, чем*); *выскоблить* (*что, на чем, чем*), *отскоблить* (*что, от чего, чем*), *проскоблить* (*что, в чем, чем*), *соскоблить* (*что, с чего, чем*); и т.п.

С учетом этого и других подобных соображений не остается сомнений в том, что глаголов с актантом Инструмент действительно намного больше, чем глаголов с актантом Средство, и что, следовательно, Инструмент имеет более высокий ранг в системе русского языка, чем Средство.

3.2.2. Невалентные свойства лексем

Помимо МУ в АС лексикографируется ряд невалентных синтаксических свойств лексем, определяющих их способность / неспособность участвовать в конструкциях широко понимаемого «малого синтаксиса». Конструкции малого синтаксиса по численности намного превосходят стандартные синтаксические конструкции языка, относительно редко встречаются в текстах и часто имеют лексически ограниченные области действия. Между тем владение ими придает речи ценное свойство идиоматичности. Приведу два примера.

Первый пример – экзистенциальные глаголы русского языка типа *бывать*, *быть*, *водиться*, *возникнуть*, *выйти*, *дуть*, *завестись*, *иметься*, *кипеть*, *найтись*, *получиться*, *произойти*, *случиться*, *стрястись*, *существовать* и т.п. Значение такого глагола, как правило, целиком включено в значение подлежащего или является его прагматической импликатурой. Иными словами, собственный семантический вклад глагола в значение предложения ничтожен. Поэтому ремой становится подлежащее: именно оно содержит максимум новой информации. В результате в нейтральных утвердительных предложениях прямой порядок подлежащего и сказуемого меняется на обратный. Ср. *В наших лесах водятся змеи*, *Возникли осложнения*, *Вышла <получилась> крупная неприятность*, *Дул сильный ветер*, *В доме завелись тараканы*, *На этот счет имеются новые данные*, *Кипят страсти*, *Нашлись дураки, которые этому поверили*, *Произошла утечка газа <перестрелка>*, *Случилось непредвиденное*, *С караульным стряслась беда*, *Существуют идеалисты, которые верят в конечное торжество добра* и т.п. Между тем для отрицательных предложений в нейтральном контексте характерен, по понятной причине, прямой порядок подлежащего и сказуемого. Ср. *Олени у нас никогда не водились*, *Осложнений не возникло*.[8]

В АС это свойство экзистенциальных глаголов единообразно, но с учетом их индивидуальных особенностей, описывается в их словарных статьях.

Второй пример – конструкция с существительным в форме ТВОР в функции обстоятельства способа, в которой существительное обозначает часть целого. Она систематически описывается в АС в словарных статьях двух групп русских глаголов, для которых она характерна.

В первую группу входят глаголы положения в пространстве, в том числе начинательные и каузативные, в

[8] Мы отвлекаемся от многих тонкостей, связанных с темо-рематическим членением высказываний, в частности, от того обстоятельства, что при определенности (известности) именной группы в позиции подлежащего восстанавливается прямой порядок слов; ср. *Раз уж такая беда стряслась с солдатом, государство обязано оплатить его лечение в любой клинике мира.* Для нас существенно лишь то, что в указанном аспекте экзистенциальные глаголы русского языка отличаются от подлинных предикатов.

004-Apresjan_J:_Layout 1 18.05.2008 14:19 Page 29

их основных значениях: *висеть (головой вниз)*, *лежать (головой к двери)*, *сидеть (боком к сцене)*, *стоять (спиной к окну)*; *ложиться (головой к двери)*, *садиться (боком к сцене)*, *становиться (спиной к окну)*; *класть (головой к двери), сажать (спиной к окну)*, *ставить (спиной к окну)*. К ним близки глаголы типа *высовываться <торчать> (кормой из воды)*, тоже в основных значениях, и выражение *выносить* (*ногами вперед*).

Во вторую группу входят глаголы пространственной ориентации типа *выходить <смотреть>* (*окнами в сад*), обычно в переносных значениях. Для них возможны трансформации вида *Дом выходит окнами в сад – Окна дома выходят в сад*, тоже фиксируемые в АС.

Обе указанные обстоятельственные конструкции следует отличать от близких по смыслу валентных конструкций с глаголами, обозначающими контакт, начало контакта или возможность контакта между двумя предметами: *доставать рукой до потолка*, *касаться рукой платья*, *опираться рукой на трость*, *прислониться спиной к двери*, *упереться лбом в стекло* и т.п., в которых существительные со значением части тела являются не только синтаксическими, но и семантическими актантами глаголов.

3.3. Сочетаемость в АС

Для активного словаря первостепенный интерес представляет, как уже было сказано выше, не вполне свободная сочетаемость слов. Ее изучение было поставлено на твердую основу в теории лексических функций (ЛФ) И.А. Мельчука; см. Мельчук 1974. Главный тезис этой теории состоит в том, что в языках мира можно выделить несколько десятков значений высокого уровня абстракции, каждое из которых выражается большим классом слов. При этом выбор конкретного слова L для выражения данного значения целиком зависит от того слова X, с которым оно сочетается. Он, тем самым, семантически не мотивирован, лексически связан, идиоматичен. Яркий пример – ЛФ MAGN = ‘большая степень того, что названо ключевым словом’. Мы говорим *кромешная тьма* и *мертвая тишина*, но не **мертвая тьма*, **кромешная тишина*; можно сказать *крепко спать* и *твердо знать*, но не **крепко знать* и **твердо спать*.

На самом деле в первых работах по ЛФ степень их идиоматичности была сильно переоценена. Главным результатом позднейших работ явился вывод, что выбор конкретного слова L на роль значения данной ЛФ от данного ключевого слова X мотивирован, хотя и не на сто процентов, каким-то общим смысловым компонентом в лексических значениях L и X (см. Апресян 2004, с дальнейшей библиографией). Наличие такого компонента объясняется общими законами семантического согласования, которое, как и всякое другое согласование в естественном языке, состоит в повторении какого-то простого элемента в составе сочетающихся слов. Это и позволяет в какой-то мере предсказывать как набор ЛФ-коллокатов для достаточно больших классов слов-аргументов, так и конкретные способы их выражения.

Посмотрим для начала на глагольную сочетаемость слова *контроль*. Вот его основные ЛФ из самого большого глагольного семейства, а именно OPER-LABOR-FUNC, включая сложные ЛФ с добавлением системных смыслов ‘начало’, ‘прекращение’, ‘каузация’ и ‘ликвидация’: *быть <находиться> под контролем*, *подвергаться контролю*, *держать кого-л. под контролем*, *подвергать что-л. контролю*, *попадать под контроль*, *выходить из-под контроля*, *ставить что-л. под контроль*, *выводить что-л. из-под контроля*.

Теперь уясним, что в значение слова *контроль* входит представление об иерархии отношений между двумя людьми или группами людей и о том, что субъект, занимающий более высокое положение в этой иерархии, может диктовать свою волю другому. В этом отношении слову *контроль* близки слова а) *власть, влияние* и б) *надзор*, *наблюдение.* Тогда, если верна гипотеза о семантическом согласовании ЛФ-глагола X с его аргументом L, естественно ожидать, что по крайней мере некоторые из этих лексически связанных глаголов будут сочетаться, помимо слова *контроль*, и с другими упомянутыми существительными. Эта гипотеза вполне оправдывается. Ср. *быть <находиться> под властью <под влиянием>*, *держать кого-л. под своей властью <под своим влиянием>, попадать под власть <под влияние>*, *выходить из-под власти <из-под влияния>, выводить кого-л. из-под власти <из-под влияния>*; *быть <находиться> под надзором <под наблюдением>*, *держать кого-л. под надзором <под наблюдением>*, *попадать под надзор <под наблюдение>*.

Иными словами, зная семантические классы ключевых слов и универсальный набор ЛФ, мы можем формировать правильные лексикографические ожидания даже по поводу не вполне свободной сочетаемости слов. Подчеркну, что сам формальный аппарат ЛФ при этом остается за кадром; он используется только для направленного сбора материала и системного представления сочетаемости слов в АС.

004-Apresjan_J:_Layout 1 18.05.2008 14:19 Page 30

3.4. Лексикализованная просодия в АС

Факты такого рода попали в поле зрения лингвистов еще в конце 80-х годов прошлого века и с тех пор изучались весьма интенсивно. В этом материале тоже были обнаружены некоторые закономерности, которые я проиллюстрирую на примере глаголов *выглядеть* и *послышаться*.

Для глагола *выглядеть* прототипическим является значение впечатления, т. е. такого образа мира, в истинности которого субъект восприятия не уверен. Ср. *Он выглядел уставшим*. Та же лексема может в определенных условиях обозначать либо факт (сдвиг в сторону объективного *быть*), либо элемент мнимого мира (сдвиг в сторону обманного *мерещиться*).

Сдвиг в сторону объективности обычно происходит во фразово безударной тематической позиции, особенно в контексте «объективных» указательных наречий *иначе*, *так*, *следующим образом* и т.п., которые берут на себя функцию ремы. Ср. *В Новой газете сообщение об этом событии выглядело иначе* [т.е. было иным]. Сдвиг в сторону мнимости возникает у *выглядеть* в рематической позиции, в частности, под логическим (контрастным) фразовым ударением. Ср. *Она только ↓↓выглядит здоровой, на самом деле она очень больной человек*.

Аналогичные сдвиги представлены на материале глагола *послышаться*,[9] с той разницей, что различия между значениями факта (*Послышалось цоканье копыт*), впечатления (*Ей послышалось, как будто в мастерской что-то тихо пробежало*) и мнимости (*– Мне послышалось, – сказал кот*) выражены гораздо более четко, так что в этом случае можно говорить о трех обособившихся лексемах. Подчеркну, что лексема *послышаться 3*, обозначающая кусочек мнимого мира, как и глагол *выглядеть* в том же значении, всегда несет главное фразовое ударение и, следовательно, рематична.

Иными словами, зная семантические классы ключевых слов и общие закономерности фразовой просодии, мы можем формировать правильные лексикографические ожидания по поводу акцентных выделений слов во фразе.

Заключение

Итак, задуман АС русского языка, учитывающий последние результаты лингвистической теории и одновременно обращенный к широкому кругу носителей языка, не имеющих специального лингвистического образования. Такой словарь должен, по замыслу, выполнять сразу две функции: а) быть компонентом полного научного описания русского языка; б) служить лексикографическим справочником активного типа. В первой ипостаси он может стать объектом разнообразных теоретических исследований. Во второй ипостаси он будет способствовать полноценному практическому овладению русским языком.

Список литературы

1. Академический словарь 1847 – Словарь церковно-славянского и русского языка, составленный вторым отделением Академии наук. СПб, 1847.

2. Апресян 1974 – Ю.Д. Апресян. Лексическая семантика. Синонимические средства языка. М., 1974.

3. Апресян 1992 – Ю.Д. Апресян. Экспериментальная, прикладная и теоретическая лингвистика: обратные связи // Festschrift für Viktor Jul'evič Rozenzvejg. Zum 80. Geburstag. Wiener Slawistischer Almanach. Sonderband 33. Wien, 1992, ss. 5-27.

4. Апресян 2003 – Ю.Д. Апресян. Системность лексики: семантические парадигмы и семантические альтернации // Etudes linguistiques romano-slaves offertes à Stanisław Karolak. Cracovie, 2003, cc. 35-47.

5. Апресян 2004 – Ю.Д. Апресян. О семантической непустоте и мотивированности глагольных лексических функций // ВЯ, 2004, № 4, 3-18.

6. Апресян 2006 – Ю.Д. Апресян. Глагольное управление revisited // Od fonemu do tekstu. Prace dedykowane Profesorowi Romanowi Laskowskiemu. Pod redakcją I. Bobrowskiego i Krystyny Kowalik. Kraków, 2006. С. 41-63.

7. Денисов, Морковкин 1978 – П.Н. Денисов, В.В. Морковкин. Учебный словарь сочетаемости слов русского языка. М., 1978.

8. Зализняк 2003 – А.А. Зализняк. Грамматический словарь русского языка. Словоизменение. М., 2003 (первое издание 1977 г.).

9. Мельчук 1974 – И.А. Мельчук. Опыт теории лингвистических моделей «Смысл Û Текст». М., 1974.

10. Мельчук, Жолковский 1984 – И.А. Мельчук и А.К. Жолковский. Толково-комбинаторный словарь современного русского языка. Вена, 1984.

[9] Впервые он был описан (независимо) в Апресян 2003: 43-44 и Падучева 2004: 200 и сл., 407.

004-Apresjan_J:_Layout 1 18.05.2008 14:19 Page 31

11. Орфоэпический словарь 1989 – Орфоэпический словарь русского языка. Произношение, ударение, грамматические формы, изд. 5-е, испр. и доп., М., 1989.

12. Падучева 2004 – Е.В. Падучева. Динамические модели в семантике лексики. М., 2004.

13. СОШ – С.И. Ожегов и Н.Ю. Шведова. Толковый словарь русского языка. М., 1992.

14. СУш – Толковый словарь русского языка. Под редакцией Д.Н. Ушакова. Тт. I – IV. М., 1934 – 1940.

15. Longman 2003 – Longman Dictionary of Contemporary English. 15-th ed. 2003.

16. Macmillan 2002 – Macmillan English Dictionary for Advanced Learners. International Student Edition. Oxford, 2002.

17. Oxford 2003 – Oxford Advanced Learner's Dictionary of Current English. 6-th edition. Oxford, 2003.

005-Ahmetova_M:_Layout 1 18.05.2008 14:28 Page 32

РЕГИОНАЛЬНАЯ ВАРИАТИВНОСТЬ ТЕРМИНОВ, СВЯЗАННЫХ С ГОРОДСКОЙ НЕДВИЖИМОСТЬЮ (ПО МАТЕРИАЛАМ ЭЛЕКТРОННОЙ БАЗЫ ПЕРИОДИКИ «ИНТЕГРУМ»)

REGIONAL VARIANTS OF THE URBAN REALTY TERMS

Ахметова М.В. *(malinxi@rambler.ru)*
Журнал «Живая старина», Москва

В докладе рассматривается не фиксируемая словарями русская региональная лексика, описывающая городское жилище, – названия типов жилых зданий и квартир (в зависимости от времени постройки, особенностей планировки, строительного материала и т.д.), термины, связанные с обозначением коммунальной квартиры и проживания в ней. В исследовании использовались материалы электронной базы периодики «Интегрум». На основе статистики упоминаний данных терминов в региональной и столичной периодике делаются предварительные выводы об ареалах распространения тех или иных номинаций.

В отличие от русских сельских диалектов, городская региональная лексика в отечественной науке практически не исследована. В немногих публикациях, появившихся за последние три десятилетия, рассматривается главным образом узус отдельных городов (например, Перми [Ерофеева 1979; Ерофеева, Скитова 1992], Томска [Банкова 1987], Элисты [Санджи-Горяева 1988], Саратова [Сиротинина 1988], Ижевска [Прокуровская 1996], Омска [Осипов 2003]); число же работ, посвященных сопоставительному рассмотрению региональной лексики, совсем невелико (например [Капанадзе, Красильникова 1982; Беликов 2004]).

С 2005 г. компания ABBYY осуществляет в Интернете проект «Языки русских городов» (www.lingvo.ru/goroda). Цель проекта – составление словаря русской региональной лексики. При проекте открыт форум «Городские диалекты», участники которого обсуждают лексические единицы, имеющие региональное распространение (на сегодня в индексе форума отмечено около 2500 «кандидатов в регионализмы»). Ареал распространения слов также проверяется по электронной базе СМИ «Интегрум» и по интернет-блогам и форумам; по результатам составляются проекты словарных статей, которые также предлагаются для обсуждения на форуме. Проект не ограничивается ни тематически, ни стилистически (обсуждаются слова от канцеляризмов до молодежного жаргона), отличается систематическим подходом (выявляются по возможности все близкие по значению слова и синонимы). Уже на данном этапе работы можно выделить некоторые тематические группы слов с наибольшим тяготением к регионализации (например, обозначения общественного транспорта, некоторых продуктов питания, видов одежды и т.д.).

Далее я обращусь к региональным обозначениям городской недвижимости, отраженные в форуме «Городские диалекты». Источником послужила периодическая печать России и ближнего зарубежья с начала 1990-х годов, размещенная в базе «Интегрум».

Вообще разговорные и региональные термины такого рода начинают попадать в словари с 1990-х годов, главным образом в специализированные (жаргонные и отражающие повседневный разговорный узус), затем и в общие толковые. Фиксировались прежде всего термины, отражающие московскую и/или петербургскую норму: *башня* [Шведова 2007], *высотка* [Corten 1992], *корабли* [Синдаловский 2002], *малогабаритка* [Corten 1992; БТС (помета «разг.»)], *малосемейка* [Corten 1992], *сталинский дом* [БТС 1998; Гусейнов 2003], *сталинские квартиры* [БТС 1998], *точечный* (о доме) [МАС 1999 (помета «спец.»), БТС 1998 (помета «спец»)], *распашонка/распашонкой* [БТС 1998 (помета «шутл.»); Елистратов 2002]. Между тем подобная лексика намного многообразнее и включает как разговорные номинации, так и канцеляризмы.

В качестве рабочего термина для обозначения рассматриваемых номинаций я буду называть их ойкемонимами (от греческого *οἶκημα* ‘жилое помещение, жилище, жилье; строение, постройка, здание’). Можно выделить несколько групп региональных ойкемонимов.

1. Ойкемонимы, мотивированные временем постройки. Наиболее продуктивная модель для образования ойкемонимов подразумевает использование суффикса *-к* в женском роде. Обычно образованное таким способом название дома переходит и на квартиру в таком доме.

005-Ahmetova_M:_Layout 1 18.05.2008 14:28 Page 33

Хрущевка в качестве названия типового пятиэтажного дома застройки 1960-х годов и квартиры в таком доме, по-видимому, универсальна.

Многоэтажное здание постройки 1930-х – первой половины 1950-х годов называется, в зависимости от региона, *сталинским домом / зданием* или *сталинкой*. Квартира в таком доме называется в первом случае *сталинской квартирой*, во второй – *сталинкой*. В *табл. 1* приведено соотношение данных терминов, наиболее показательные цифры выделены «маркером»; курсивом выделены названия городов, относительно которых на основании имеющихся сведений сделать выводы сложно. Москва и Петербург, где распространена номинация *сталинский дом / здание (квартира)* (в их орбиту включены близкие города, например Ярославль для Москвы и Калининград для Петербурга), противопоставлены остальной России, где преобладают *сталинки*, показатели на *сталинский дом* малы, а на *сталинское здание* и *квартиру* – ничтожны. Выделяются Приуралье, Урал и частично Западная Сибирь, где показатели на *сталинский дом* по отношению к *сталинке* довольно высоки.

	Сталинка	**Сталинский дом**	**Сталинское здание**	**Сталинская квартира**
ЗАРУБЕЖЬЕ				
Украина	446	52	6	4
Беларусь	76	32	3	13
Латвия	179	119		1
СЕВЕРО-ЗАПАД				
Петербург	353	428	22	39
Калининград	7	20		
Петрозаводск	2	7		
Сыктывкар	2	1		1
Архангельская обл., Псков, Новгород, Мурманск	212	46	1	3
ЦЕНТР				
Москва	1161	1745	108	138
Московская обл.	21	20		
Ярославль	5	14	1	
Тула	24	19	2	1
Курск	6	2		
Белгород	8	2		
остальной Центр	645	48	2	4
ЮГ	368	34	3	3
ПРИВОЛЖЬЕ				
Уфа, Киров, Ижевск, Чебоксары, Йошкар-Ола	27	24	1	2
остальное Приволжье	822	52	2	5
УРАЛ				
Екатеринбург	18	9	1	3
Пермская обл.	37	13		2
Челябинская обл.	46	29	1	7
Оренбург	30	1		1
Курган	4	5		
СИБИРЬ				
Тюмень	45	2	2	
Тюменская обл.	5	3		
Омск	5	4		
Томск, Кемеровская обл., Новосибирск	60	31	2	4
Барнаул, Абакан, Красноярский край, Иркутск, Чита	263	15	1	1
Бурятия, Якутия	3	4	1	
ДАЛЬНИЙ ВОСТОК	259	14		6

Таблица 1.

005-Ahmetova_M:_Layout 1 18.05.2008 14:28 Page 34

Большие показатели на *сталинку* в столицах следует трактовать как влияние регионов. Пока опрошенные мной жители Москвы и Петербурга не признают *сталинку* характерным для своих городов словом. В *табл. 2* показана динамика изменения в прессе столиц соотношения *сталинка / сталинский дом* по годам, в сравнении с несколькими газетами из городов, дающих наиболее высокие показатели по *сталинкам* (Ростов-на-Дону, Нижний Новгород, Самара и Саратов). В прессе столиц *сталинка* начала входить в широкое употребление лишь в 2004 – 2005 гг., а в 2006 г. наметилось преобладание *сталинок* в московских газетах, посвященных недвижимости, т.е. можно говорить о вытеснении *сталинками сталинских домов* в московском риэлтерском сленге.

Петербург (все издания)	Москва (все издания)	«Газета Дона» (Ростов-на-Дону)	«Проспект» (Нижний Новгород)	«Пульс Поволжья» (Самара)	«Саратовский Арбат»
	до 1993: 0/12				
1995: 0/5	1994—95: 2/26				
1996—97 1/14	1996—97: 12/96				
1998—99 4/10	1998—99: 26/171	до 1999: 14/0	до 1999: 17/0	до 1999: 13/0	до 1999: 21/2
2000—01 14/36	2000—01: 68/196				
2002—03 35/65	2002—03: 155/336				
2004: 81/74	2004: 109/172	2000—04: 27/2	2000—04: 87/2	2000—04: 10/0	2000—04: 70/0
2005: 82/96	2005: 165/226				
2006: 125/131	2006: 293/177	2005—06: 9/0	2005—06: 3/1	2005—06: 4/0	2005—06: 6/0

Таблица 2.

В ряде регионов типовые многоэтажные дома постройки 1960 – 1980-х годов носят название *брежневка* (Украина, Беларусь, Латвия, Волхов, Северодвинск, Калининград, Центральная Россия исключая Орел, Курск, Липецк, Брянск и Калужскую обл., Ростовская обл., Среднее Поволжье, Урал без Оренбурга, Тюмень, Хабаровск и Сахалин); в столицах эта номинация известна риэлтерам. Требует уточнения не только ареал *брежневки*, но и ее значение, поскольку, согласно одним текстам, они отличаются от хрущевок улучшенной планировкой и имеют 5 этажей; согласно другим, этажей в них 9.

Петербургскую *кировку* (здание постройки начала 1930-х годов) следует также отнести к сленгу (встречается только в риэлтерских текстах, опрошенные петербуржцы слова не слышали).

2. Ойкемонимы, обусловленные функциональными характеристиками зданий. Слово *малосемейный* в словарях интерпретируется как ‘имеющий небольшую семью’. Видимо, формула *жилье (общежитие, дом) для малосемейных* трансформировалась в *малосемейное жилье (общежитие, дом)*, которое сократилось до *малосемейки*, а слово *малосемейный* приобрело значение, не фиксируемое словарями, – ‘рассчитанный на небольшую семью’.

Можно выделить следующие значения термина *малосемейка*:

а) ‘общежитие для малосемейных’:

(1) *Четыре человека погибли во время ночного пожара в малосемейке обанкротившегося завода химволокна в Костанае* («Казахстанская правда», Алматы; 18.12.1998);

б) ‘жилой дом с малогабаритными квартирами’:

(2) *В этот же период появились так называемые малосемейки – дома коридорного типа с однокомнатными квартирами* («Биржа плюс свой дом» (Н.Новгород); 26.04.2001);

в) ‘малогабаритная квартира в жилом доме или общежитии’:

(3) *Как матери-одиночке Ирине выдали «малосемейку», однокомнатную квартирку в общежитии* («Орловский Меридиан»; 05.04.2000);

Малосемейка распространена практически везде, кроме Москвы, Тулы, Петербурга, Пскова, Карелии, Калмыкии, Северного Кавказа без Адыгеи. Нуждаются в уточнении Курск, Воронеж, Среднее Поволжье, Восточная Сибирь и Дальний Восток.

Другим ойкемонимом, ареал которого довольно широк и исключает столицы, является *гостинка.* При этом исходная формула *дом/квартира/комната гостиничного типа* известна в Москве. *Гостинки* локализуются в следующих ареалах: Украина; Эстония; Северо-Запад (Мурманская обл., Новгород, Коряжма Архангельской обл., Калининградская обл.); Кострома, Рязань; Ростовская обл., Астрахань; Приволжье без Мордовии, Татарстана, Пензы и Башкирии; Урал (Березники Пермской обл. и Курган); Западная и Центральная Сибирь, исключая тюменский север; Владивосток и Хабаровск. Показательно, что *гостинки* отсутствуют в Петербурге,

005-Ahmetova_M:_Layout 1 18.05.2008 14:28 Page 35

Уфе, Архангельске и Оренбурге, т.е. в городах, где существует соответствующий топоним (от *Гостиный двор*).

По той же модели образованы *улучшенка* ‘дом или квартира улучшенной планировки’ (Ростовская обл.; Кострома, Иваново, Рязань; частично Приволжье и Урал; Красноярский край; Хабаровск), *малогабаритка* ‘малогабаритная квартира’ (Москва), *благоустройка* ‘благоустроенный дом’ (Архангельская обл. и Бурятия), *секционка* ‘жилье в общежитии секционного типа’ (Чувашия, Курганская обл., Томск, Кемеровская обл., Красноярск).

3. Ойкемонимы, отражающие обеспеченность коммунальными удобствами. В *табл. 3* приведено распределение в периодике терминов *с частичными удобствами* (ЧУ), *полублагоустроенное* (ПБ) и *частично благоустроенное* (ЧБ):

	ЧУ	ПБ	ЧБ		ЧУ	ПБ	ЧБ
ЦЕНТР				**ПРИВОЛЖЬЕ**			
Москва	90	12	9	Саров Нижегородской обл.			6
Тула	1			Йошкар-Ола			19
остальной Центр	705	2	7	Киров	4	2	6
ЗАРУБЕЖЬЕ				*Ижевск*		3	
Украина	53			остальное Приволжье	277	1	5
Донецк	10		5	**УРАЛ**			
Беларусь	62		21	Оренбург	22		
Латвия	336			Пермская, Свердловская обл., Курган		73	1
Молдова	9			*Челябинская обл.*		1	
СЕВЕРО-ЗАПАД				**СИБИРЬ**			
Петербург, Луга, Выборг	9	2	2	Тюмень	2	37	
Тихвин, Волхов	27			*Тюменская обл.*	3		2
Калиниградская обл., Псков, Мурманск, Вологда, Новгород	295	3	7	Омск, Томск, Кемеровская обл., Новосибирск	3	130	2
Петрозаводск	1	54	3	*Красноярск, Абакан, Горно-Алтайск*	1	6	2
Архангельск	34	2	2	Алтайский край			11
Архангельская обл.	1		37	Кызыл, Иркутск, Чита, Улан-Удэ	2	54	
Нарьян-Мар	34			Якутия		9	72
Сыктывкар	1	6	10	**ДАЛЬНИЙ ВОСТОК**			
ЮГ И КАВКАЗ	737	1		Хабаровск	43	2	4
				Биробиджан	6		1
				остальной Дальний Восток	2	8	3

Таблица 3.

Интересно, что внутри одного региона и даже области могут бытовать разные термины (ЧУ в Архангельске и ЧБ в Архангельской обл.; ЧУ в Оренбургской обл., ПБ в соседних Пермской, Свердловской и Курганской и полное отсутствие подобных формулировок в Челябинской обл.). Есть и «белые пятна» (хотя едва ли такое жилье в этих регионах отсутствует), например Москва: для нее 90 упоминаний в прессе слишком незначительно.

Иногда от данных терминов образуются разговорные номинации, в данном случае благодаря сокращениям в газетных объявлениях. В Карелии от ПБ образуется *полублаг*; в Якутии от ЧБ образуется *чэбэшка*; в Краснодарском крае отмечена *частичка* (в последнем случае можно говорить о тройной мотивации: часть частного дома с частичными удобствами).

4. Коммунальные квартиры. Номинации *коммунальная квартира* и разговорное *коммуналка* универсальны. Но есть альтернативные региональные обозначения как самой коммунальной квартиры, так и способов проживания в ней. Это обозначения, связанные с понятиями *общая кухня* и *подселение*.

Общая кухня в качестве самостоятельной номинации, обозначающей квартиру на несколько квартиросъемщиков, практически не встречается. Используются выражения: *жить на общей кухне* ‘жить в коммунальной квартире’; *комната на общей кухне* ‘комната в коммунальной квартире’; *квартира на общей кухне* ‘коммунальная квартира’.

005-Ahmetova_M:_Layout 1 18.05.2008 14:28 Page 36

Подселение используется самостоятельно в значении ‘доля коммунальной квартиры, занимаемая отдельным жильцом или семьей’:

(4) *Меняю 1 подселение* [указан адрес] *на подселение в любом р-не* («Новое Омское Слово»; 21.07.1999).

Форма *на общей кухне* распространена в Беларуси, Калининградской обл., Брянске, Пензе, Саранске и Кургане. *Подселение* с формами *жить на подселении* и *жить с подселением* известно шире; следует отметить на Севере Архангельскую обл. и Вологду, в Центре – Калужскую, Тульскую и Липецкую обл., на юге – Волгоградскую обл., на Урале – Пермскую, Челябинскую и Оренбургскую обл.; Западную Сибирь до Алтая, Красноярский край; Владивосток и Хабаровск.

5. Названия квартир, мотивируемые количеством комнат, обычно совпадают с региональным обозначением числа как условного номера: ‘однокомнатная квартира’ – *однерка* (Пермь)*, однешка* (Кемеровская обл., Томск, Новосибирск). *Полуторка* представляет собой межрегиональный омоним: в Челябинске это ‘крупногабаритная однокомнатная квартира’, в других регионах – ‘малогабаритная двухкомнатная квартира’ (Белоруссия, Латвия, Юг, Пенза, Екатеринбург, Западная Сибирь без Новосибирска, Иркутск, Дальний Восток, возможно Псков, Коми и Вологда).

6. Ойкемонимы, образованные от строительных проектов и серий: *грузинка* (от *грузинский проект*, Крым), *ленинградка* (от *ленинградская серия*, Татарстан, Самарская обл., Удмуртия, Челябинская обл., Омск, Кемеровская обл., Красноярский край), *литовка* (Латвия, Мордовия), *ульяновка* (Архангельская обл.), *чешка* (Украина, Белоруссия, Молдова, Крым, Ростовская обл., Воронеж), *киевка* (Тюмень, Чебоксары, Новороссийск, Черновцы), *ленпроект* и *моспроект* (Ханты-Мансийский АО, *капэдэшка* (от серии КПД 464 ВМ, Якутия), *спецпроект* (от *специальный проект*, Латвия, Украина, Крым, Екатеринбург).

7. Ойкемонимы, мотивированные внешним видом дома: а) особая форма – *трехлистник* (Ханты-Мансийский АО, Ноябрьск), *каскадник* (Рыбинск Ярославской обл.); б) многоэтажный одноподъездный дом: *башня* (Москва), *точка* (Петербург), *шишка* (Троицк Московской обл.), *свечка* (Киев, Обнинск, Ярославль, Владимир, Волгоград, Ростовская обл., Краснодар, Среднее Поволжье, часть Урала, Сибирь до Красноярска, Владивосток); в) протяженность – *многоквартирник, двухквартирник, трехквартирник* и т.д. (Урал и Сибирь).

8. Ойкемонимы, мотивированные материалом: *деревяшка* ‘многоквартирный деревянный дом; квартира в таком доме’ (Северо-Запад без Петербурга и Калининграда, Приуралье, Пермская обл., Сибирь, Якутия, Приамурье), *капиталка* ‘панельный дом; квартира в таком доме’ (север Тюменской обл.).

9. Обозначения планировки квартиры: а) со смежными комнатами – *вагончиком* (четкая регионализация не выявлена)*, икарусом* (Хабаровск)*, трамвайчик / трамвайчиком* (Донбасс, Одесса, Ростовская обл., Краснодарский край, Самарская обл., Нижний Новгород, Пермская обл., Челябинск, Оренбургская обл., Кемеровская обл., Новосибирск, Хакасия), *смежка* (Одесская обл); б) планировка, при которой комнаты расположены по обе стороны коридора: *распашонка* (повсеместно)*, бабочка* (Ростовская обл.)*, самолет* (Тольятти); в молдавских газетах несколько раз встречается термин *рубашка*, но пока нет уверенности, что это не окказионализм.

10. Среди **ойкемонимов-историзмов**, запечатлевших ушедшие в прошлое реалии,можно назвать нижегородскую *народную стройку* ‘малоэтажный жилой дом; квартира в таком доме’, является сокращением от *дом народной стройки* и отсылает к соответствующей строительской практике:

(5) *...Снесут так называемые «народные стройки» – малоэтажные дома, которые в 60-х годах возводили сами обитатели* («Биржа» (Н.Новгород); 08.12.2003).

На Юге (Краснодарский край, Адыгея, Таганрог Ростовской обл., возможно Пятигорск) малоэтажный многоквартирный дом называется *жактовским домом*, или *жактом*. Отличительной особенностью жактов является общий, или жактовский, двор, который служит семантической заменой собственно дома, например:

(6) *В жактовских дворах живут по 10 – 12 семей* («Краснодарские известия»; 04.07.2000).

11. Обозначения временного жилья могут применяться в значениях, не совпадающих со словарными.

Так, *пансионат*, интерпретируемый словарями как род гостиницы или дома отдыха, в Тюмени и Сургуте имеет значения ‘общежитие квартирного типа; квартира в таком общежитии’:

(7) *Но это правило не распространяется на общежития квартирного типа, или, как их еще называют, малосемейки, пансионаты* («Тюменский курьер»; 27.05.2000);

Привычные толкования слова *коттедж* подразумевают благоустроенность этого типа жилья («небольшой жилой благоустроенный дом в пригороде, в (рабочем) поселке» [Ожегов 1986]; «1. небольшой жилой дом в пригороде, в рабочем поселке и т. п. для одной семьи, обычно двухэтажный (второй этаж – мансарда). 2. легкая (обычно летняя) постройка в туристских лагерях» [МАС 1999]; «индивидуальный городской или сельский жилой дом (обычно двухэтажный) с небольшим участком земли» [БТС 1998]). В современном узусе наряду с универсальным значением, относящимся к индивидуальному, часто элитному жилью существует и региональное – ‘временное или предполагающееся как временное жилье’, распространенное на севере

005-Ahmetova_M:_Layout 1 18.05.2008 14:28 Page 37

Тюменской обл.:

(8) *И только в 1967 году мы въехали в настоящую квартиру – выделили в первых одноэтажных деревянных домах с громким именем «коттедж»* («Нефть Приобья» (Сургут); 11.09.1999).

Впрочем, уже в 1960 – 1970-е годы благоустроенность коттеджей была относительной, не говоря уже о современности:

(9) *Жили вдевятером в коттедже, где было всего шесть кроватей…*» («Слово Нефтяника» (Ноябрьск Ямало-Ненецкий АО); 23.05.2003);

(10) *Поселили временно снова в деревянный коттедж, только похуже, без санузла с ванной, с прогнившим полом и дырявым потолком* («Сургутская Трибуна»; 14.08.2002).

12. Экспрессивные ойкемонимы *молодоженка* ‘семейное общежитие’ (Липецк), *хихишник* ‘дом гостиничного типа с дешевыми квартирами’ (Владивосток).

Таким образом, базы СМИ могут служить важным источником для установления региональной дистрибуции лексических единиц. Однако, используя одни лишь базы, установить ареал того или иного слова возможно не всегда, и требуются дополнительные средства, помогающие уточнить этот ареал (проверка региональной литературы, блогов и форумов, интервью и т.д.).

Некоторые лексические единицы могут быть адекватно описаны по базе СМИ; так, слово *чэбэшка* с орфографическими вариантами встретилось в пяти газетах (30 текстов), причем только из Якутии, что позволяет подготовить полноценную словарную статью:

ЧЭБЭ′ШКА, Якутия. *Разг*. Орф. вар.: **чебэ́шка**, реже **чебе́шка** (от сокр. *ч/б* ‘частично благоустроенный’). 1. Дом, не полностью обеспеченный коммунальными удобствами. *С давних пор ее двухкомнатная квартирка в чэбэшке стала местом боевых сражений* (Якутия; 16.06.2006). 2. Квартира в таком доме. *На 2007 год городской округ выделил из муниципального бюджета поселку 600 тысяч рублей (цена однокомнатной «чебэшки»)* (Якутия; 13.02.2007).

В других случаях (особенно для жаргонной лексики) уточнение семантики и ареала бытования слова требует обращения к блогам, форумам и непосредственной работы с информантами.

На примере семантического поля «городская недвижимость» мы можем видеть, что специфическая для конкретного региона лексика формируется независимо, хотя и может испытывать влияние со стороны соседних регионов. Поскольку она не является единой для всего русскоязычного пространства (даже если она называет не узкорегиональные, а общекультурные реалии), налицо актуальность исследования и фиксации региональных форм городской лексики.

Список литературы

1. Банкова 1987 – Банкова Т.Б. Лексика томского городского просторечия: Дис. ... канд. филол. наук. Томск, 1987.

2. Беликов 2004 – Беликов В.И. Сравнение Петербурга с Москвой и другие соображения по социальной лексикографии // Русский язык сегодня. Вып. 3: Проблемы русской лексикографии. М.: Ин-т рус. яз. РАН, 2004. С. 23 – 38.

3. БТС 1998 – Большой толковый словарь русского языка / [Сост., гл. ред. С.А. Кузнецов]. СПб.: Норинт, 1998.

4. Гусейнов 2003 – Гусейнов Г.Ч. Д.С.П.: Материалы к Русскому Словарю общественно-политического языка конца XX века. М.: Три квадрата, 2003.

5. Елистратов 2002 – Елистратов В.С. Словарь русского арго: Электрон. версия. 2002. [Электрон. ресурс: www.slovari.gramota.ru/elistratov].

6. Ерофеева 1979 – Ерофеева Т.И. Локальная окрашенность литературной разговорной речи: Учеб. пособие. Пермь: Изд-во Перм. гос. ун-та, 1979.

7. Ерофеева, Скитова 1992 – Ерофеева Т.И., Скитова Ф.Л. Локализмы в литературной речи горожан. Пермь: Изд-во Пермского гос. ун-та, 1992.

8. Капанадзе, Красильникова 1982 – Капанадзе Л.А., Красильникова Е.В. Лексика города (к постановке проблемы) // Способы номинации в современном русском языке / Отв. ред. Д.Н. Шмелев. М.: Наука, 1982. С. 282 – 294.

9. МАС 1999 – Словарь русского языка: В 4 т. / Под ред. А.П. Евгеньевой. (Малый академический словарь). М. Рус. яз., 1999.

10. Ожегов 1986 – Ожегов С.И. Словарь русского языка / Под ред. Н.Ю. Шведовой. М.: Рус. яз., 1986.

11. Осипов 2003 – Словарь современного русского города / Под ред. Б.И. Осипова. М.: Русские словари; АСТ; Астрель; Транзиткнига, 2003.

005-Ahmetova_M:_Layout 1 18.05.2008 14:28 Page 38

12. Прокуровская 1996 – Прокуровская Н.А. Город в зеркале своего языка. Ижевск: Изд-во Удмурт. ун-та, 1996.

13. Санджи-Горяева 1988 – Санджи-Горяева З.С. Некоторые особенности устной речи г. Элисты // Разновидности городской устной речи. М.: Наука, 1988.

14. Синдаловский 2002 – Синдаловский Н.А. Словарь петербуржца. СПб.: Норинт, 2002.

15. Сиротинина 1988 – Сиротинина О.Б. Языковой облик г. Саратова // Разновидности городской устной речи. М.: Наука, 1988. С. 247 – 252.

16. Шведова 2007 – Толковый словарь русского языка с включением сведений о происхождении слов / Отв. ред. Н.Ю. Шведова. М.: Издательский центр «Азбуковник», 2007.

17. Corten 1992 – Corten I.H. Vocabulary of Soviet society and culture: a selected guide to Russian words, idioms, and expressions of the post-Stalin era, 1953 – 1991. London, 1992.

006-Baranov_A:_Layout 1 18.05.2008 14:20 Page 39

ПРОТИВ «РАЗЛОЖЕНИЯ СМЫСЛА»: УЗНАВАНИЕ В СЕМАНТИКЕ ИДИОМ

AGAINST DECOMPOSITION OF MEANING: RECOGNITION IN SEMANTICS OF IDIOMS

Баранов А.Н. *(baranov_anatoly@hotmail.com)*
Институт русского языка им. В.В.Виноградова РАН

В докладе обсуждается проблема отражения в модели значения (в частности в толковании) образной составляющей семантики идиом. Предлагается для передачи внутренней формы использовать не атомистическую стратегию разложения смысла на более простые составляющие, а подход, основанный на узнавании и реконструкции образа по семантическому триггеру в толковании, инициирующему цепочку ассоциаций, которые порождают нужный образ.

1. Атомистический подход к значению

Основания современной лексической (шире – лингвистической) семантики, интересовавшие исследователей в шестидесятых-семидесятых годах XX в., дискуссии о семантических метаязыках описания за редкими исключениями канули в лету. Семантическая «технология» и неопозитивистские идеалы научности явственно уступают свои позиции изучению языкового факта, поддержанному теоретическими и практическими исследованиями функционирования языка в дискурсе, а также широким распространением когнитивной методологии, понимаемой зачастую чрезвычайно широко и неопределенно. Если и появляются новые идеи общетеоретического и методологического характера, то они носят явно вторичный характер к описанию языкового материала – ср., например, идею семантических «кварков», меньших, чем семантические примитивы, высказанную Ю.Д. Апресяном [Апресян 2004], или семантических «фотонов», сформулированную Е.В. Урысон [Урысон 2004]. Эти теоретические «бантики» - блики на фоне многоводной реки лингвистического факта – только подчеркивают красоту самой реки, которой нельзя не восхищаться.

Возникает ощущение достижения конечной истины в теоретических представлениях о сущности значения и семантики языковых форм. Выделение очередных «сверхбесплотных» и «серхминимальных» единиц семантического метаязыка только убеждает в окончательной победе атомистического подхода к исследованию плана содержания языковых выражений. Похоже, работа над созданием семантических метаязыков в русле исследований по машинному переводу, начатая в СССР 60-х гг. XX в., успешно завершена, по крайней мере, в России и, по крайней мере, в теоретическом аспекте.

Доструктуралистсткие теории о существовании связей между означающим знака и его планом содержания преодолены. Конечно, идея иконичности блуждает мелким соблазном (см. работы А.Е. Кибрика и Дж. Хэймана), да и критика Р. Якобсоном концепции произвольности знака Ф. де Соссюра имеет место, однако в целом атомистический подход к описанию семантики признается правильным и единственно научным.

В то же время именно обращение к языковому факту – к материи языка – порождает многочисленные вопросы к атомистической идее описания плана содержания, в основе которой лежит стратегия разложения интерпретируемого выражения на более простые – в идеале минимальные – смысловые единицы. Свидетельств недостаточности атомистического подхода накопилось довольно много. Основная сложность заключается в экспликации плана содержания недискретных феноменов языка – метафор (и значений, образованных на основе тропеических преобразований), слов с процедурной семантикой (в частности частиц), фразеологических единиц, актуальное значение которых возникло как результат переосмысления внутренней формы. Здесь я рассмотрю лишь одну из черных дыр атомистического подхода к значению – семантику идиом[1].

[1] О системно обусловленных проблемах атомистического подхода к значению см., в частности, [Баранов, Кобозева 1987], [Baranov 2007].

006-Baranov_A:_Layout 1 18.05.2008 14:20 Page 40

2. Семантические «триггеры» узнавания образа

Значение идиом с живой внутренней формой мотивировано образом, лежащим в ее основе. Например, идиомы *прикинуться ветошью* и *прикинуться шлангом* очень близки и по значению, однако их образы все-таки несколько различаются, и это оказывается достаточным для дифференциации их актуальных значений: метафора ветоши высвечивает в семантике идиомы идею незначительности и незаметности, которая, в свою очередь, вызывает к жизни идею непричастности к чему-то происходящему. Метафора шланга профилирует представление о сущности, легко меняющей свое положение под действием внешних сил и неспособной к мышлению. Отсюда и актуальные значения этих выражений:

прикинуться ветошью = ‘пытаться своим поведением создать у других участников ситуации представление о собственной незначительности и непричастности к происходящему, чтобы избежать неприятных для себя последствий’;

прикинуться шлангом = ‘пытаться своим поведением создать у других участников ситуации представление о собственном непонимании чего-то очевидного и при этом нежелательного для субъекта’.

Типовая ситуация употребления идиомы *прикинуться ветошью* – это описание попытки уйти от опасности. Ср. *На всякий случай охранник упал, отполз в кусты и* ***прикинулся ветошью***. В этом контексте идиома *прикинуться шлангом* явно неуместна: ???*На всякий случай охранник упал, отполз в кусты и* ***прикинулся шлангом.***

Идиома *взлететь на воздух* в значении ‘быть разрушенным или уничтоженным в результате воздействия каких-л. взрывчатых веществ’ недопустима при описании террористического акта: взрыва самолета в воздухе (не дай бог!), поскольку во внутренней форме есть связь с «воздухом». При этом взрыв самолета на стоянке этой идиомой описать можно, хотя это и оказывается на грани нормы и приобретает уже черты игрового употребления: *Один из самолетов на стоянке при попытке разминирования взлетел на воздух*.

Совершенно прозрачное влияние внутренней формы на актуальное значение обнаруживается у пар идиом *помалкивать/молчать в тряпочку* и *помалкивать/молчать в портянку*[2]. Обе эти идиомы описывают ситуацию вынужденного молчания человека из-за опасения возникновения каких-то проблем. Однако, если первая идиома относительно универсальна, то вторая используется, как правило, для описания ситуаций, связанных с отношениями между военнослужащими: - *Так блин, служивые, закрыть хлебало и* ***молчать в портянку****. Пока вы два года дурью маялись, мы на вас тут пахали.* В обычном контексте она выглядит несколько странно: *<…> какие же мы маленькие <…>. В прямом смысле слова. Самый рослый из нас, Пашка, на метр от земли возвышается, остальные вовсе воробьи. И сверху вниз взирают на карликов великаны. И карлики* ***молчат в тряпочку/***???***портянку****, подавленные своей ничтожностью и бессилием.* [Ю. Нагибин. Бунташный остров]. Часть образа «портянка» недвусмысленно отсылает к реалиям армейского быта, причем живость образа ограничивает сферу употребления идиомы.

Подобные примеры легко умножить[3]. Насколько аналитическая часть толкования эксплицирует вклад образной составляющей идиомы в ее семантику? В какой-то степени это можно сделать, отразив те семантические следствия, которые профилируются соответствующей метафорой: в приведенных выше моделях значения[4] идиом *прикинуться ветошью* и *прикинуться шлангом* воспроизводятся следствия из метафоры ветоши как чего-то ненужного и незначительного (‘представление о собственной незначительности и непричастности к происходящему’) метафоры шланга как чего-то податливого и лишенного даже минимальных признаков мышления (‘представление о собственном непонимании чего-то очевидного’). Что касается метафор тряпочки и портянки в идиомах *помалкивать/молчать в тряпочку* и *помалкивать/молчать в портянку*, то с этим совсем плохо. Актуальные значения указанных единиц совпадают: ‘не высказывать своего мнения по какому-л. поводу, поскольку это может повлечь неприятные последствия’, а образ в виде следствия из метафоры здесь отразить трудно. Конечно, ограничение на употребление идиомы *помалкивать/молчать в портянку* можно грубо ввести в форме ‘используется для описания ситуаций, связанных с отношениями между военнослужащими’, но понятно, что и по стилю, и по сути это, скорее, комментарий к толкованию, чем само толкование. Между тем, образ, метафора, лежащая в приведенных примерах идиом за актуальным значением, явственно ощущаются носителями языка. Это проявляется, например, в целенаправленном обыгрывании этих образов, в частности в материализации соответствующих метафор. Ср. – ***Возьми****, голубь,* ***тряпицу*** *и* ***молчи*** *в нее!*; - *Солдат, как печенье? Добавить зубной пасты? Не вопрос! А ты, солобон, рта не раскрывай.* ***Молчи в портянку****!* ***Своей не хватает, вот тебе моя****!*

---

[2] Толкования данных идиом – Баранов А.Н., Добровольский Д.О.

[3] Целый ряд примеров такого рода приводится в [Баранов, Добровольский 1998].

[4] Здесь и далее термин «модель значения» используется как более общая категория экспликации семантики языкового выражения по сравнению с толкованием. См. по этому поводу [Баранов 1996; Баранов, Добровольский 2008].

006-Baranov_A:_Layout 1 18.05.2008 14:20 Page 41

Описание метафоры во внутренней форме едва ли можно свести к аналитическому описанию ее области источника и области цели. Метафора должна быть представлена в модели значения как нечто цельное – единый образ, целостный гештальт, неразложимый на отдельные составляющие. Для этого естественно обратиться к другой стратегии толкования, основанной не на разложении смысла, а на его **узнавании** как концепта, уже имеющегося в сознании носителя языка. Именно в этом случае появляется возможность апеллировать к метафоре, к образу, как к единому гештальту. Такая стратегия толкования должна основываться не на аналитическом описании метафоры, а на намеке, на «триггере», спускающем крючок ассоциаций, приводящих к нужному образу. Здесь уместно предложить аналогию угадывания слова в кроссворде: именно так может быть устроен триггер-намек. В случае идиомы *прикинуться шлангом* метафору можно отразить в деепричастном обороте, вводимом оператором ‘как бы’, который переводит последующую часть семантической экспликации в возможный мир: ‘пытаться своим поведением создать у других участников ситуации представление о собственном непонимании чего-то очевидного и при этом нежелательного для субъекта, как бы выдавая себя за неодушевленный объект, априори не обладающий способностью к мышлению и лишенный собственной воли’. Здесь намек обеспечивается родовым обозначением (‘неодушевленный объект’) и указанием свойств, профилирующих один из коммуникативных центров актуального значения – ‘непонимание субъектом чего-то очевидного’. Шланг, конечно же, «не понимает», но «способен сделать» так, как скажут, поскольку не имеет собственной воли. Отсюда слабое семантическое следствие притворной готовности изменить поведение. Это следствие, однако, слишком слабое, чтобы быть эксплицитно представленным в толковании, хотя оно относительно легко реконструируется в результате сопоставления собственно аналитической части толкования и «компонента узнавания» - семантического триггера, передающего образ идиомы.

Родовой триггер, хотя и более высокой степени абстракции, лежит в основе представления образа и в идиоме *прикинуться ветошью* = ‘пытаться своим поведением создать у других участников ситуации представление о собственной незначительности и непричастности к происходящему, чтобы избежать неприятных для себя последствий, как бы выдавая себя за что-то неодушевленное – незаметное и никому не нужное’.

В самом простом случае семантический триггер образа может быть упрощенным толкованием слов, передающих образ во внутренней форме. Ср. модель значения идиомы *обивать пороги*: ‘много раз безуспешно приходить в какое-л. учреждение с просьбой о чем-л., что осмысляется как многократное вхождение в помещение, сопровождаемое задеванием за нижнюю часть дверной коробки’. По этому же принципу организован триггер в идиоме *будь друг(ом)*: ‘выражение настоятельной просьбы собеседнику сделать что-л. в форме предложения начать испытывать дружеские чувства к просящему’. Казалось бы, использование аналитического толкования для семантического намека на образ противоречит исходной посылке об использовании стратегии узнавания смысла, однако это не так. Узнавание образа происходит в этом случае по его следам – компонентам семантического разложения, а «толчок» к узнаванию дает соответствующий оператор, вводящий семантический триггер – ‘осмысляемый как’, ‘ассоциируется’, ‘уподобляется’, ‘в форме/имеющий форму’, ‘как бы + деепричастный оборот’, ‘описывается’, ‘сопоставляется’ (об операторах такого типа в модели значения идиомы см. подробнее [Баранов, Добровольский 1998; 2008]). Использование родового триггера (см. выше случаи идиом типа *прикинуться ветошью*) по сравнению с более или менее подробным толкованием является более сложным намеком на образ, поскольку данных для реконструкции образа родовой триггер несет существенно меньше.

Семантический триггер, ведущий к распознаванию образа, может быть и совершенно иной природы. В тех случаях, когда образ в идиоме мотивирован прецедентным текстом или культурной реалии, триггер может иметь форму отсылки к соответствующему культурному феномену, ср. модель значения идиомы *бальзаковского возраста*, в которой семантический триггер вводится оператором ‘осмысляемый как’: ‘женщина среднего возраста или несколько старше, сохранившая привлекательность, демонстрирующая интерес к мужчинам и пользующаяся у них успехом, осмысляемая как женский персонаж произведений О. де Бальзака’. Аналогично, хотя и с другим оператором, представлен семантический триггер образа в значениях идиомы *богатенький Буратино*: 1) ‘разбогатевший человек, имеющий возможность неразумно тратить деньги на свои прихоти, уподобляемый известному непослушному герою сказки А.Н. Толстого, часто совершавшему ошибочные поступки’; 2) ‘богатый человек как объект возможных действий по отъему денег или другого ресурса, уподобляемый известному герою сказки А.Н. Толстого, обманутому разбойниками’. Отметим, что два значения идиомы *богатенький Буратино* появляются в результате коммуникативного высвечивания различных сторон персонажа-источника: «свойственное Буратино непослушание vs. Буратино как жертва обмана».

Внутренняя форма многих идиом абсурдна в том смысле, что она основывается на случайном фонетическом сходстве двух слов, одно из которых соответствует какому-то компоненту актуального значения, а второе – кладется в основу образа. Семантический триггер толкования должен включать указание на это

006-Baranov_A:_Layout 1 18.05.2008 14:20 Page 42

фонетическое сходство. Ср. *завтраками кормить (кого-л.)*: ‘регулярно обещать кому-л. сделать что-л. в ближайшем будущем – чаще на следующий день – и не выполнять обещанного, что осмысляется как регулярно повторяющаяся неуместная попытка обещающего организовать утренний прием пищи, причем такое осмысление мотивировано фонетическим сходством слов «завтрак» и «завтра», последнее из которых намекает на типичную попытку отказаться от выполнения обещания, перенеся свою деятельность по выполнению обещанного в будущее’. Ср. также толкование речевой формулы *[- Ну?] – Баранки гну!*: ‘оценка неуместности речевого поведения собеседника как ответная реакция на его реплику, содержащую частицу «ну» и настоятельно и, тем самым, невежливо побуждающую говорящего сделать что-л., в форме указания на собственную занятость абсурдным действием придания круглой формы одному из видов хлебобулочных изделий, уже имеющему круглую форму, причем выбор названия действия определяется исключительно тем, что оно рифмуется с первой репликой собеседника и тем самым утрированно имитирует его речевое поведение’. Имитация фонетической основы внутренней формы может относиться не только к фонетическому сходству, но и к особенностям говорения в той или иной ситуации. Ср. *Хватай мешки/узлы – вокзал отходит*: ‘указание на необходимость быстро и не раздумывая сделать что-л. из-за опасности опоздания и возникновения из-за этого серьезных проблем, выраженное в форме призыва, быстро взяв вещи, немедленно садиться на поезд, причем в призыве, имитирующем спешку и волнение говорящего, слово «поезд» перепутано со словом, обозначающим место прихода и отправления поездов’.

Перечислить все типы семантических триггеров, создающих эффект узнавания образа, не представляется возможным просто потому, что описание фразеологии в этом направлении только началось[5]. В определенной степени они связаны с типом оператора, вводящим семантических триггер. Для правильного выбора и формулирования семантического триггера важным оказывается и сам образ – модель внутренней формы, чаще всего это – метафора, лежащая в основе актуального значения идиомы.

3. Модели внутренней формы идиомы

Способ указания на актуальное значение, фиксированный во внутренней форме идиомы, можно назвать **моделью внутренней формы**. Значительная часть моделей внутренней формы основана на метонимии, понимаемой в широком смысле. В **метонимических моделях** во внутренней форме фиксируется часть процедуры, действия, общей ситуации, хотя имеется в виду целое – вся процедура, все действие, вся ситуация. Указание в метонимических моделях может основываться на описании невербального поведения, часто представленного в ритуализованных процедурах выражения соответствующей коммуникативной интенции (в том числе на жесте): *поклониться в ноги/ножки*[6]*; памятник поставить; бить поклоны; склонить голову; [взять] под козырёк; ударить по рукам; по рукам; щёлкать каблуками; повернуться лицом (к кому-л./чему-л.); во фрунт*. Метонимическими можно считать и такие модели, в которых фиксируются характерные признаки ситуации, связанной с выражаемой коммуникативной интенции, ср. *семь футов [воды] под килем; барабан на шею, флаг в руки.* Метонимическими по природе являются модели, основанные на псевдоисчерпании[7], ср. *ни сват ни брат*; *ни кола ни двора, чтоб … стоял и деньги были.* К метонимическим моделям относятся также такие способы указания на актуальное значение, которые предполагают выбор одного из элементов некоторой последовательности или процедуры: *до гробовой доски, по гроб жизни, до гроба/могилы*, *с/от младых/молодых ногтей, со школьной скамьи*.

Синонимическая модель характеризуется тем, что в качестве указания на актуальное значение используются два квазисинонима: *целиком и полностью*, *любо-дорого смотреть/глядеть, чин-чинарём, чин-чином.*

Довольно часто во внутренней форме идиом встречается **модель множества**. Она часто реализуется как отрицание наличия даже одного элемента некоторого множества: *ни звука, ни пылинки/соринки, ни копейки/гроша, ни души, ни капли, ни грамма, ни крошки, ни грамма, ни грана, ни на йоту*. Отрицание может отсутствовать, что приводит к модификации актуального значения в противоположную сторону – *до копейки*. Модель множества может реализоваться и в ровно противоположной стратегии, когда во внутренней форме указывается на все множество (*всем миром*); все множество и каждый из его элементов (*всем и каждому; все и каждый; ты да я, да мы с тобой*); все множество и отдельный его элемент (*все до последнего, все до единого*) или несколько разных множеств (*все/всё и вся*). Множество может задаваться последовательным перечислением

[5] Первый словарь идиом, передающий в толкованиях образный компонент на основе стратегии узнавания, уже создан, но пока не опубликован, см. [Баранов, Вознесенская, Добровольский, Киселева, Козеренко в печати]. Некоторые приводимые здесь толкования идиом использованы в этом словаре.

[6] Валентности идиом здесь и далее для простоты опущены.

[7] О псевдоисчерпании как характеристике идиоматичности см. [Баранов, Добровольский 2008].

006-Baranov_A:_Layout 1 18.05.2008 14:20 Page 43

(*по капле, шаг за шагом, капля за каплей, день ото дня, изо дня в день*), представляться формулой, полностью исчерпывающей его содержание – **модель взаимодополняющих подмножеств** (*душой и телом*), множество может указываться по своим крайним членам (*от А до Я, альфа и омега, [и] стар и млад, от мала до велика, [и] нашим и вашим*). С моделью множества тесно связана **модель счета** элементов множества: *потерять счёт, для ровного/круглого счёта, сбиться со счёта/счёту, ровным счётом, не в счёт, сбрасывать со счёта/счетов, поставить на счётчик; на раз, на счёт раз; в два счёта*.

Модель пространства близка модели множества в том смысле, что пространство можно осмыслять как множество точек на поверхности. Модель пространства представлена такими частными моделями, как **иерархическое пространство** (*на край света, край земли; на краю земли; на край земли*), **неиерархическое пространство** (*вдоль и поперёк, от края [и] до края*), **отрицание конечной точки/области пространства** (*без конца и без края, конца-края нет, ни конца ни края*). **Модели времени** также концептуально близки модели множества, поскольку моменты времени естественно осмысляются как элементы упорядоченного множества (кортежа). Ср., например, указание на начальный и конечный момент некоторой временной последовательности (*от зари до зари; от темна до темна; с утра до вечера/ночи*), описание взаимодополняющих интервалов общего временного интервала (*денно и нощно, день и ночь, днями и ночами, дневать и ночевать*).

Образная часть семантики идиомы часто основывается на сравнении. **Модель эксплицитного сравнения** широко представлена в русской идиоматике. Ср. идиомы *как снежный ком, как по команде, как с куста, [как] по мановению волшебной палочки, [как] по мановению руки, лететь... как на крыльях, нестись… как сумасшедший, бежать как крысы с [тонущего] корабля, стоять/сидеть... как истукан ми т.д.* Эта модель может использоваться для указания на самые различные смыслы – ‘быстро’ (*как из пушки, как штык*), ‘много/тесно’ (*как сельди в бочке; как кильки в банке/бочке*), ‘медленно’ (*ползти… как черепаха*), ‘неожиданно’ (*как чёртик/чёрт из табакерки*), ‘окончательно/полностью’ (*разойтись как в море корабли*). Внутри этой модели можно выделить **модель притворного сравнения**, которая часто реализуется в идиомах для кодировки смысла ‘ненужности’. Ср. *[нужен] как зайцу стоп-сигнал/бубен/триппер/модная болезнь/колокольчик...; [нужен] как козе баян; нужен как пятое колесо в телеге; нужен как попу гармонь; нужен как рыбе зонтик* и т.д.

В русской идиоматике представлены и другие модели внутренней формы. В данном случае существенно, что модель указания на актуальное значение может рассматриваться как важная эвристика для выбора семантического триггера, включающего цепочку ассоциаций, которые порождают нужный образ. Например, **модель семантически немотивированного фонетического уподобления** основана на случайном фонетическом сходстве лексического компонента реплики собеседника и компонента ответа говорящего. Она используется для указания на неуместность речевого акта собеседника. Ср. речевые формулы *[- Ну?] – Дышло гну; [- Кто?] – Дед пихто; [- Почему?] – По кочану; [- Где?] – В Караганде; [- Кто?] – Конь в пальто; [- Говорят.] – [Где-то] кур доят*. Намек на образ в толковании таких форм должен во всех случаях содержать указание на фонетическое сходство. См. выше толкование речевой формулы *[- Ну?] – Баранки гну!*. Ср. также *[- Где?] – В Караганде*: ‘оценка как неуместного вопроса собеседника о местонахождении кого-л./чего-л. в форме ответа, указывающего на город, название которого рифмуется с вопросительным словом вопроса собеседника «где», а также с конечной частью нецензурного слова, часто используемого в ответной реплике с похожим значением’. Последний компонент толкования отражает стилистический статус данной речевой формулы как эвфемизма.

В семантических триггерах толкований идиом *дырка от бублика* и *от жилетки рукава [получить...]* (**модель несуществующей части объекта**) должно указываться, что в метафоре происходит сравнение с чем-то несуществующим[8]:

дырка от бублика = ‘полное отсутствие ресурса как результат его несправедливого распределения, сопоставляемое с несъедобной – и более того, **нематериальной** – частью хлебобулочного изделия, при том что другим досталась съедобная’;

от жилетки рукава [получить...] = ‘не получить ничего от ожидавшегося ресурса из-за его несправедливого распределения, что осмысляется как получение **несуществующей** части чего-л.’

4. Заключение

Узнавание смысла по «намеку» в модели значения (как техника толкования) вряд ли может использоваться в словарях, предназначенных для обучения языку. Этот способ экспликации смысла подойдет только носителям языка, погруженным в соответствующий языковой и культурный контекст. Но это одновременно отражает и объективные сложности в освоении фразеологии неносителем данного языка. Очевидно, что идиоматика усваивается существенно позже на достаточно солидном базисе более простых языковых форм. Стратегия

[8] Толкования данных идиом – Баранов А.Н., Добровольский Д.О.

006-Baranov_A:_Layout 1 18.05.2008 14:20 Page 44

узнавания как способ экспликации семантики применима только к таким языковым выражениям, актуальное значение которых возникает на основе метафорической переинтерпретации, к «недискретной семантике», так или иначе связанной с образом.

Данная работа, разумеется, полемична. Я ни в коей мере не выступаю за полную аннигиляцию атомистического подхода к плану содержания языковых выражений с торжественным выносом и погребением его теоретических положений. Меня смущает, во-первых, широкое обсуждение теоретических основ семантики – в том числе семантики естественного языка – в современной логической и философской литературе (в этих многочисленных феноменологических, ситуационных, прагматических и пр. семантических теориях[9]) при явно недостаточном внимании к основаниям семантической теории в отечественной лингвистике. Во-вторых, мне кажется чрезвычайно важным рассмотрение сложностей экспликации семантики «недискретных семантических феноменов» (к которым относится метафора и, конечно, образная часть плана содержания фразеологизмов) при использовании имеющихся семантических метаязыков. Наконец, в-третьих, применение атомистических техник к анализу идиом, предполагающих разложение целого на простейшие семантические компоненты, ассоциируется у меня с взрывом семантики идиомы с последующей попыткой установить по уцелевшим кускам, как она устроена.

Список литературы

1. Апресян Ю.Д. Акциональность и стативность как сокровенные смыслы (охота на оказывать) // Сокровенные смыслы. М., 2004.

2. Баранов А.Н. Служебные слова как объект исследования авторской лексикографии (по крайней мере vs. по меньшей мере в художественных текстах Ф.М. Достоевского) // Слово Достоевского. М., 1996

3. Баранов А.Н., Вознесенская М.М., Добровольский Д.О., Киселева К.Л., Козеренко А.Д. Малый фразеологический словарь: Значение, образ, употребление // под ред. А.Н. Баранова, Д.О. Добровольского. (в печати)

4. Баранов А.Н., Добровольский Д.О. Внутренняя форма и проблема толкования // Изв. РАН. СЛЯ. Т.57, № 1, 1998. С. 36-44.

5. Баранов А.Н., Добровольский Д.О. Аспекты теории фразеологии. М., 2008 [в печати].

6. Баранов А.Н., Кобозева И.М. Метаязыковые средства описания семантики предложения: опыт типологии // Лингвистическое обеспечение информационных систем. М., 1987.

7. Урысон Е.В. Некоторые значения союза А в свете современной семантической теории // Русский язык в научном освещении. 2004. № 2.

8. Baranov A.N. The problem of the metalanguage // Phraseologie: ein internationals Handbuch zeitgenössischer Forschung / ed. by H. Burger [et al.]. Walter de Gruyter, Berlin, N.-Y., 2007. P. 77-90.

9. Dummett M. The Seas of Language. Oxford: Clarendon Press, 1993.

10. Dennett D. Evolution, Error and Intentionaly // Sourcebook on the Foundations of Artificial Intelligence, in Y. Wilks and D. Partridge, eds, New Mexico University Press, 1988.

[9] Достаточно посмотреть лишь пару работ в этой области – например, [Dummett 1993] или [Dennett 1988], чтобы понять, что поднимаемые там вопросы имеют прямое отношение к лингвистической семантике.

007-Belikov_VI:_Layout 1 18.05.2008 14:39 Page 45

ПАРЕМИИ КАК ОБЪЕКТ ЛЕКСИКОГРАФИИ
LEXICOGRAPHY OF PROVERBS

Беликов В.И. *(vibelikov@mtu-net.ru)*
Институт русского языка им. В.В.Виноградова РАН

В докладе рассматривается практика создания словников, фиксирующих наиболее употребительные русские пословицы и поговорки. Показано, что и состав таких словников, и выбор основного варианта конкретных паремий субъективен. Предлагаются пути решения встающих в этой области проблем.

В перечне сложностей толковой лексикографии на первых местах окажутся выделение отдельных значений лексической единицы, аккуратность толкований, стилистическая квалификация единицы в целом или ее конкретного значения, подбор иллюстраций и т. п. – все эти проблемы кроются внутри словарной статьи. Перед составителем фразеологического (в широком смысле) словаря встает другая общеизвестная проблема: упорядочивание самих словарных статей. При этом необходимо соблюсти баланс интересов: адресат должен легко находить в словаре нужную информацию, но и составитель хочет хоть как-то отразить собственный взгляд на организацию фразеологического (паремиологического) фонда в расстановке единиц в словаре. Принципы устройства словаря сообщаются в предисловии, но редкий читатель согласен учиться пользованию книгой и знакомиться с предисловием. Так что предельно толерантный к потребителю своей продукции паремиолог, вроде бы, должен поместить пословицу[1] везде, где читателю вздумается ее искать. Такой подход, кажется, никогда не применялся, поскольку однотомник стал бы многотомником, с пропорциональным ростом цены издания и времени поиска, – а в этом как раз не заинтересован именно потребитель. Промежуточное решение – организовать словарь по некоему принципу (например, алфавитному), снабдив его указателем.

Именно таким образом организован словарь В.П. Жукова [1998], где в указатель включены абсолютно все входящие в паремии слова; в результате слова *баба* отсылает к шести пословицам, на союз *а* отведено четыре страницы петита, на отрицание *не* – почти восемь. Конечно, объемный сборник так организовать трудно, но в этот включено лишь «1200 пословиц и поговорок, наиболее часто употребляемых в русской речи и зафиксированных в письменных литературных источниках» [Жуков 1998: 4]. Пробуем искать «общеизвестные» пословицы *Работа не волк, в лес не убежит*, *Поперёд батьки в пекло не суйся* и *Вилами по воде писано*, но на своем алфавитном месте их нет. В первом случае читатель вряд ли будет продолжать искать пословицу, во втором, вероятно, попытается найти замену начальному *поперёд*, но тоже безуспешно; однако старательная работа с Указателем наверняка поможет, поскольку все три пословицы в словаре Жукова представлены: *Дело не медведь* (*не волк*), *в лес не уйдет* (*не убежит*); *Прежде отца* (*прежде батьки*) *в петлю не суйся* (*не лезь*); *Это <еще> вилами на воде писано*.

«Как показывают факты, многие пословицы и поговорки бытуют в разных вариантах ‹…› не всегда легко отграничить общенародный вариант от индивидуально-авторского» [Жуков 1998: 15]. Заголовком словарной статьи служит «типичная, наиболее употребительная форма» паремии [Жуков 1998: 26], в круглых скобках приводятся варианты – лексические, как в приведенных выше пословицах, формальные – *Волков* (*волка*) *бояться – в лес не ходить*, словообразовательные – *Своя рубашка* (*рубаха*) *ближе к телу*, структурные – *Видна птица по полету* («исходная форма»), *Видно птицу по полету* («структурный вариант»). «Существенно отличным» от вариативности считается явление факультативности; заключаемая в словаре Жукова в угловые скобки факультативная часть может оказаться по сути служебным словом (как в примере выше, однако замечу, что местоимение *это*, по Жукову, является непременной, неопускаемой частью пословицы), а может быть содержательной – *Вольному воля <спасенному рай>*.

О заменяемых и опускаемых фрагментах паремий специалисты упоминали постоянно, но с таким же постоянством говорилось об их клишированности, неизменности; не случайно основная работа главного паремиолога-теоретика имеет подзаголовок «Заметки по общей теории клише» [Пермяков 1970].

[1] Или поговорку. Я отдаю себе отчет в различии пословиц и поговорок, в том, что паремии не сводятся к этим названным, а также в том, что при определенном взгляде на фразеологию никакие паремии в нее не включаются. Строго говоря, речь идет о «паремиях пословичного типа» (включая, скажем, прецедентные тексты), как раз в этом смысле ниже я и использую без различения термины *паремия* и *пословица*, относя их к фразеологии.

007-Belikov_VI:_Layout 1 18.05.2008 14:39 Page 46

В действительности особого противоречия тут нет: любая собственно пословица (предложение) или поговорка (словосочетание) в речевых произведениях, конечно, изменчива, но оказывается невероятно клишированной при сопоставлении с аналогичными «обычными» текстовыми единицами. Так что речевой вариативностью паремий поначалу просто пренебрегали, как трением при выявлении основных законов механики.

Вариативность, отражаемая словарями пословиц (и свойственная отнюдь не всем паремиям), – это вариативность не текстовая, а системная; эта вариативность не отменяет их клишированности, так же, как варьирование типа *феноме́н* / *фено́мен* (или *инцидент*/ *инциндент*) не позволяет усомниться в единстве соответствующих лексических единиц в рамках литературной нормы (или общенародного языка).

Однако если вариативность лексической единицы – нечастое исключение, в паремиологическом фонде это норма. Пока пословицы оставались в ведении филологии (фольклористики и «филологизированного» языкознания), проблема вариативности просто не вставала. Семиотический подход к паремиям начинается с работ Г.Л. Пермякова, которого интересовала в первую очередь типология паремиологических клише. Вероятно, первой работой, специально посвященной вариативности пословиц, стала статья Е.Н. Саввиной [1984], но она касалась речевых трансформаций. В последние годы исследования в этой области активизировались, но они также либо охватывают лишь текстовую вариативность (например, Жигарина 2006), либо ведутся теми, кто не видит нужды в четком противопоставлении языка и речи. Вариативность пословиц как единиц словаря специально не исследуется. А без этого невозможно выделение прототипических паремий и отделение их парадигматики (системного варьирования) от синтагматики (текстового варьирования). То есть то, что предоставляют нам словники пословиц – собрание ценного, но случайного материала.

В самом деле, из рассмотренных выше трех пословиц две вошли в известный «Паремиологический минимум» Пермякова [1988]: *Работа не волк* (*не медведь*)*: в лес не убежит*; *Наперед батьки в петлю* (*в пекло*) *не суйся*. Жуков для этих пословиц предложил по четыре варианта, Пермяков по два, каждый выбрал основной; совпадений вообще нет. Аналогию представляла бы пара двуязычных словарей, где неким русским словам давалось бы по два—четыре перевода, но всегда разных.

А как «на самом деле»? В НКРЯ *дело не медведь* встречается только у Фета, Лескова, Чехова (дважды), *дело не волк* – у Салтыкова-Щедрина и Мельникова-Печерского. *Работа не медведь* – один раз у Мамина-Сибиряка, *работа не волк* встретилась в 9 документах, самое раннее вхождение – в кинофильме «Операция «Ы» и другие приключения Шурика». Похоже, Жуков отражает норму XIX века, Пермяков – современную, но для достоверного вывода данных мало. Воспользуемся базой СМИ «Интегрум». *Работа не волк* в русскоязычных бумажных СМИ фигурирует более 1400 раз[2], остальные вместе взятые – 20 (*дело не медведь* и *дело не волк* – по пять, *работа не медведь* – десять). «Не медведь» в текстах XIX в. пять раз «не уходит», дважды «не убегает»; в современной периодике глаголы при *медведе* распределяются поровну (4:4); с *волком* же сочетаемость глагола *уйти* пренебрежимо ничтожна: шесть примеров при 640 «не убежит».

У второй пословицы «Интегрум» нашел всего четыре варианта с *петлей*: в «Правде КПРФ» (*вперед батьки в петлю*) и русскоязычных газетах Украины (*поспешили наперед отца в петлю*; *«видите, что получается, когда прежде отца в петлю...»*; *не спешит сунуться впереди отца в петлю*); основных вариантов Жукова и Пермякова нет вообще. *Отец* и *пекло* в бумажных СМИ в рамках этой паремии не встретились, но дважды нашлись в материалах украинских информагентств, в вариантах *поперед отца в пекло* и *впереди отца в пекло*[3]. Этим единичным примерам противопоставлены варианты с *батькой* и *пеклом*, различающиеся начальным предлогом: *поперед*: 224, *поперек*: 230, *вперед*: 121, *наперед*: 17, *впереди*: 11, *прежде*: 6, *раньше*: 4 и несколько единичных вариантов (*попередь батьки*, *перед батькой* и др.). Равно- (и высоко-) вероятными зачинами ядра пословицы являются предлоги *поперед* и *поперек*, которым заметно уступает *вперед*; вместе они обслуживают 94% вхождений пословицы в СМИ[4]. Эта общая картина меняется в русскоязычной прессе Украины: в 18 из 22 такого рода примеров использован предлог *поперед*.

В третьей из упомянутых пословиц также проявляется непредусмотренное Жуковым варьирование. В НКРЯ 9 раз встретилось *вилами по воде* и 6 раз *вилами на воде*. В современных СМИ России в целом отмечается незначительное преобладание варианта *по воде* (351 против 343 с предлогом *на* в региональной прессе, 212:185 в центральных газетах), хотя по регионам соотношение оказывается различным (Петербург, 23 издания – 29:7, Якутск, 4 изд. – 1:10, Казань, 5 изд. – 3:11); для русскоязычной прессы Украины, Белоруссии, Молдавии

[2] Почти в половине случаев без продолжения. Иногда встречаются авторские переработки, как правило, меняющие смысл пословицы: *…утром не загрызет; убежит – не поймаешь; но она иногда убегает; но убежать может; но от человека убегает; но что скажет волк-работодатель; но и на нее есть охотники; работа – ворк* [=work]*; поймать ее трудно, но охотников хватает* и т. п.

[3] В этой украинской по происхождению пословице (наиболее частотный вариант – *не лізь поперед батька в пекло*) *отец* заменяет *батьку* только в текстах с Украины; очевидная гиперкоррекция.

[4] Детали о предикате (чаще предшествующем ядру) опускаю, но *не лезь* наиболее частотно, *не суйся* уступает ему более, чем на порядок.

007-Belikov_VI:_Layout 1 18.05.2008 14:39 Page 47

характерно стойкое преобладание варианта с *по* (153:3, 16:1, 12:0). Зачин *Это <еще>* вряд ли следует считать частью пословицы, он выполняет дейктическую функцию при включении пословицы в ткань текста. Нередко подлежащее при этом обороте достаточно абстрактно (чаще всего – *всё это*, *это* но также *и то и другое*, *остальное* и т. п.), но эта позиция может быть заполнена и предложением (*как на нас там смотрят, еще вилами писано по воде* – Данилевский, Княжна Тараканова), и именем (с которым согласуется глагол): *Эта дата была писана вилами на воде* (Суханов, Записки о революции); *Надежды ‹…› на скорый прием в Североатлантический блок вилами по воде писаны* (МК, 1.04.08).

Важной характеристикой паремии является степень ее распространенности. Попытка объективировать этот показатель предпринималась лишь однажды, когда в начале 1970-х гг. на основе представительного анкетировании был подготовлен «Паремиологический минимум». Минимум оправданно стал основой нескольких двуязычных словарей пословиц советского периода, продолжает использоваться с этой целью до сих пор и по-прежнему квалифицируется как собрание паремий, «которые являются общеизвестными и активно употребляются русскоговорящими» [Rodak 2002: 48]. Не вдаваясь в подробности подготовки минимума, отмечу, что в него попали не все «общеизвестные и активно употребляющиеся» пословицы; кроме того, общеизвестность и активное использование коррелируют не всегда.

Общеизвестность может быть результатом чтения классической литературы и современной периодики, так что одни всем известные паремии архаичны, другие стали всего лишь газетными штампами. Широта использования в устной и письменной речи может различаться на два порядка, а во втором случае сильно зависеть от жанра; уровень известности устаревающих и вновь распространяющихся пословиц серьезно различается по возрастам. Приблизиться к объективному учету всех этих обстоятельств стало возможным лишь с появлением больших массивов оцифрованных текстов.

Собственно устный узус статистическому исследованию еще недоступен, но отдаленной его моделью могут служить некоторые типы интернет-коммуникации. Пока поисковые машины не позволяют выявлять статистику только в непринужденных типах общения, но некоторым приближением к таким данным служит результат поиска по блогам. Можно сопоставить число вхождений определенных паремий (фрагментов паремий) в базу «Интегрум» (бумажные СМИ) и результаты поиска Яндекса по блогам[5] (Табл. 1).

Поисковый образ*	бумажные СМИ	Яndex-блоги	соотношение
1. "работа не волк"	1437	13095	0,1
2. "без мыла" влезть/лезть	183 (106/77)	1061 (574/487)	0,2
3. !груши околачивать	344	1713	0,2
4. "на халяву"/"на холяву"/нахаляву !уксус	427	1531	0,3
5. "быстро только кошки" [!родятся]	101 [81]	352 [232]	0,3 [0,3]
6. "кто перв(ый,-ым)/раньше встал, того и тапки"	490 (378/112)	1583 (1774/809)	0,310
7. "вилами по/на воде "	1513 (907/606)	4221 (3127/1094)	0,4
8. "при ловле блох" [!спешка !нужна] (!нужна/!хороша); (!спешка/!поспешность)**	486 [151] (240/151); (271/73)	785 [305] (436/167); (505/60)	0,62 [0,49]
9. "голод не тетка" [!пирожка]	1901 [29]	2865 [28]	0,66
10. "овчинка выделки не стоит"	1515	1985	0,8
11. "готовь сани летом" [!телегу]	4691 [1105]	5270 [223]	0,9 [5,0]
12. обжечься "на молоке" ["на воду"]	1901 [1565]	2030 [1006]	0,9 [1,6]
13. "по Сеньке [и] шапка"	1598	1637	1,0
14. "была бы шея" [!хомут]	143 [81]	141 [87]	1,0 [0,9]
15. "куй железо, пока горячо"	629	603	1,0
16. "поперед/поперек/вперед батьки" !пекло	575 (224/230/121)	524 (133/193/198)	1,1
17. "не все коту масленица" ["великий пост"]	2530 [233]	854 [36]	3,0 [6,5]
18. "в чужом пиру похмелье"	291	84	3,5
19. "стыд не дым" [глаз]	221 [187]	52 [37]	4,3 [5,1]
20. "сколько веревочке ни/не виться"*** "как веревочка ни/не вейся" "сколько веревочку/веревку ни/не вить"	3093 (2355/738) 47/13 10/1	534 (124/410) 6/21 4/7	5,8 2,2 1,0
21. "по заслугам [и] честь"	906	43	21,0

Таблица 1.

[5] Следует иметь в виду, что при поиске Яндексом первое выданное число вхождений некоего текста может отличаться от конечного результата в разы; недавно введенное по примеру Гугла ограничение числа показываемых сайтов (по блогам – страниц) делает числа более 1000 малодостоверными; чем больше число, тем сильнее оно завышено.

007-Belikov_VI:_Layout 1 18.05.2008 14:39 Page 48

Примечания. * В моей нотации кавычки, их отсутствие и «!» имеют те же значения, что в Яндексе, через дробь обозначены альтернативные варианты, в квадратных скобках – опускаемая часть; после суммарной цифры в скобках иногда приведены данные по вариантам. Паремии **11**, **18**, **21** отсутствуют у Жукова (Ж), №№ **7**, **13**, **19** – у Пермякова (П), №№ **2**—**6** и **8** – в обоих сборниках. Пословицы №№ **1**, **7**, **16** подробно рассмотрены выше, остальные представлены в следующем виде (у Ж. при двух вариантах, оформленных отдельными предложениями, первым идет основной): **9**: *Голод не тетка <пирожка не подсунет>* (Ж), *Голод не тетка* (П); **10**: *Овчинка выделки не стоит*; *Овчинка выделки стоит* (Ж), *Овчинка выделки не стоит* (П); **11**: *Готовь сани летом, а телегу – зимой* (П); **12**: *Обжегся на молоке, дует и на воду*; *Обжегшись на молоке, станешь* (*будешь*) *дуть и на воду* (Ж), *Обжегшись на молоке, дуют* [*и*] *на воду* (П); **13**: *По Сеньке <и> шапка <по Ерёме колпак* (*кафтан*)*>* (Ж); **14**: *Была бы шея, <а> хомут найдется* (Ж), *Была бы шея, а хомут найдется* (П); **15**: *Куй железо, пока горячо* (Ж, П); **17**: *Не все коту масленица <бывает и великий пост>* (Ж), *Не все коту масленица* [*— будет и великий пост*] (П); **18**: *В чужом пиру похмелье* (П); **19**: *Стыд не дым, глаза не выест* (*не ест*) (Ж); **20**: *Сколько веревку* (*веревочку*) *ни вить, а концу быть*; *Как веревочка ни вейся, а концу быть* (Ж), *Сколько веревочке ни виться, а концу быть* (П); **21**: *По заслугам и честь* (П).

** Пословица 8 бытует в большом числе частотных вариантов; ее ядро (*при ловле блох*) встретилось в 486 текстах СМИ, чаще всего в сочетании со *спешка нужна* (151 текст). Наиболее частое подлежащее в этой пословице – *спешка* (271 текст), сказуемое – *нужна* (240 текстов), но корреляции между ними нет.

*** Как видим, четверть журналистов и три четверти блоггеров ошибаются при написании отрицания не/ни, что имеет важные следствия для поиска.

Строки таблицы упорядочены по соотношению вхождений паремии в бумажные СМИ и интернет-блоги. Для того, чтобы интерпретировать этот коэффициент как показатель различий во встречаемости пословиц в соответствующих типах текстов, следует знать, как соотносится число статей бумажных СМИ и страниц блогов, какова частотность «обычных» слов в текстах двух жанров. Получить убедительные результаты в этой области достаточно трудно. Для грубой оценки я просчитал вхождение в два типа текстов отдельных словоформ из исследуемых паремий (Табл. 2). Общие цифры по всем массивам были бы очень велики, а по блогам совершенно недостоверны, поэтому я сравнил тексты за два периода 2007 г.

словоформа	янв.—март 2007		сент. 2007	
	СМИ	блоги	СМИ	блоги
"+не стоит"	25052	161314	7865	29584
!зимой	17475	80409	5077	13523
!волк	1915	17264	484	6633
!тапки	186	13210	60	2791
!шея	647	10971	191	2669
!чужом	1339	7155	497	2355
!уксус	499	1638	259	639
!готовь	280	1283	131	487
!ловле	416	968	152	179
!веревочке	282	735	80	259
!виться	253	457	82	184
!обжегшись	88	272	33	92

Таблица 2.

В прессе отражается сезонное колебание тематики: о волках чаще пишут в зимние месяцы, а об уксусе – в период заготовок; сезонные колебания в блогах не всегда объяснимы (а четырех- и особенно пятизначные числа поисковая машина сильно завышает). Но кое-что это сопоставление дает. В периодике относительно редко употребляются бытовые слова *тапки* и *шея*[6], напротив, *ловля* и *виться* – слова «неблоговые». Не будет серьезной ошибкой считать, что на начало 2008 г. частота появления нейтральных лексических единиц в блогах была в 3—4 раза выше, чем в рассматриваемой базе периодики.

Теперь можно грубо соотнести частоту встречаемости отдельных паремий: в повседневном бытовом дискурсе *Работа не волк* используется в десятки раз чаще, чем в прессе, а *По заслугам и честь* – в десятки раз реже. Важны, конечно, и абсолютные цифры, невысокие цифры в блогах обычно являются свидетельством книжности, а совсем малые цифры в обоих источниках (полный вариант № 9, основной вариант Жукова № 20) – свидетельство заведомой устарелости (и в прессе, и в Интернете практически все вхождения этих вариантов – цитирование пословичных сборников).

[6] Слово *шея*, конечно, стилистически нейтрально, но в быту оно употребляется заведомо чаще, чем в газетах.

007-Belikov_VI:_Layout 1 18.05.2008 14:39 Page 49

Объем доклада не позволяет детально остановиться на многих частных выводах, но главный очевиден: состав обоих сборников субъективен. Более половины паремий Табл. 1 не включены в одно или оба собрания. В ряде случаев причины понятны. №№ 4 и 6 получили широкое распространение после выхода обеих публикаций (1 изд. словаря Жукова – 1966 г.). №№ 2, 3 и 5 в некоторых своих вариантах обсценны. Но взаимные противоречия двух словников и неполное соответствие реальному словоупотреблению и без того многочисленны.

Словарная фиксация откровенно обсценных паремий – отдельная проблема. Но есть такие, которые достаточно широко используются в двух вариантах, обсценном и необсценном, внешне различаясь либо степенью допустимости используемых вариантов, либо наличием/отсутствием неудобопроизносимого фрагмента. Происхождение их различно. Иногда обсценный вариант заведомо вторичен и представляет собой результат дисфемизирующей языковой игры; такие варианты редко становятся широкоизвестными. Другой путь появления подобной пары – сознательная эвфемизация сниженного в коммуникативном контексте, для которого первичный вариант расценивается как недопустимый. Такого рода пословиц, способных к «мимикрии», довольно много; когда-то я назвал их паремиями-трикстерами [Беликов 1994: 256]. Трикстерный их характер проявляется в том, что часть носителей русского языка, зная их лишь в «переодетом» виде, не накладывает ограничений на контексты, где они могут быть использованы.

В ходе обсуждения подобных паремий на форуме «Охота за цитатами» (http://forum.lingvo.ru/actualtopics.aspx?bid=29) из случайной опечатки для обозначения «благопристойного» варианта паремии-трикстера возник вполне удачный термин эвфеминизм. Эвфеминизмы нередко проникают в лексикографические издания как единственные представители соответствующих пословиц; по крайней мере в ряде случаев это следствие ущербности знаний лексикографа в области описываемого паремиологического фонда. Так, из паремий Табл. 1 в иллюстративных примерах БТС не представлены лишь №№ 1, 5, 6, 19 и 21; в частности, в статьях **влезть** и **груша** фигурируют варианты *Без мыла влезет кто-л. куда-л.* (без помет) и *Груши* (*с дерева*) *околачивать* (с пометой *разг.-сниж.*). Обе пословицы довольно старые, обсценные их варианты имеются среди не вошедших в основной корпус материалов В. И.Даля: *Без мыла в жопу лезет*; *Хуем груши околачивать* [Русские… 1997: 487, 505].

Резюмировать сказанное можно следующим образом. Идеальный словарь паремий в качестве инварианта должен использовать синхронно преобладающий (в необходимых случаях – равноправные по встречаемости); иные варианты (в том числе прототипические с исторической точки зрения) должны присутствовать в словарной статье с соответствующими стилистическими и обобщенно-частотными пометами. При этом необходимы сведения о территориальных и социальных различиях в статусе вариантов паремии (объем не позволил детально остановиться на этих аспектах). «Грамматика паремии», способы реализации ее в тексте, должны быть описаны, но синтагматическим вариантам в словарной статье не место.

Список литературы

1. Беликов В.И. Паремиологические заметки // Знак: Сборник статей по лингвистике, семиотике и поэтике памяти А.Н. Журинского – М.: Русский учебный центр, 1994.

2. Жигарина Е.Е. Современное бытование пословиц: вариативность и полифункциональность текстов. АКД. М., 2006.

3. Жуков В.П. Словарь русских пословиц и поговорок. – 6-е изд. – М: Рус. яз., 1998.

4. Пермяков Г.Л. 500 наиболее употребительных русских пословичных изречений // Пермяков Г.Л. Основы структурной паремиологии / Сост. Г.Л. Капчиц – М.: ГРВЛ изд. «Наука», 1988.

5. Пермяков Г.Л. От поговорки до сказки (Заметки по общей теории клише). М: ГРВЛ изд. «Наука», 1970.

6. Русские заветные пословицы и поговорки (В.И. Даля) // Народные русские сказки не для печати, заветные пословицы и поговорки, собранные и обработанные А.Н. Афанасьевым. 1857—1862. М., «Ладомир», 1997.

7. Саввина Е.Н. О трансформациях клишированных выражений в речи // Паремиологические исследования: Сб. ст. / Сост., ред. и предисл. Г.Л. Пермякова – М., 1984.

8. Rodak Anna. Заметки о новом переводном словаре русских пословиц // Semiosis lexicographica, XI, Warszawa, 2002. S. 48.

008-Bogdanov_A:_Layout 1 18.05.2008 14:21 Page 50

ОРФОГРАФИЯ В ИНТЕРНЕТЕ: АНАЛИЗ ОДНОЙ ОРФОГРАФИЧЕСКОЙ ОШИБКИ

ORTHOGRAPHY IN THE INTERNET: THE ANALYSIS OF ONE MISSPELL

Богданов А.В. *(bidon@inbox.ru)*
Московский государственный университет им. М.В. Ломоносова

В работе рассматривается орфография в сети Интернет как естественный языковой процесс. Делается анализ одной, отдельно взятой, орфографической ошибки, а именно ошибочного написания мягкого знака в формах третьего лица единственного числа возвратных глаголов *(делается, придеться)*. Далее делается попытка доказать, что орфографические ошибки могут различаться по своей частотности, а значит и по своему праву на установление новой орфографической нормы.

1. Введение

С развитием Интернета, в том числе и русскоязычной его части, у людей, активно пользующихся сетью, появилось новое языковое пространство – в этом пространстве, с одной стороны, приходится использовать письменный язык, но с другой – нет практически никаких ограничений на правильность его использования. В самом деле, до появления Интернета практически все сферы использования письменного языка были таковы, что в них не допускались или, по крайней мере, не приветствовались орфографические ошибки. Будь то диктант по русскому языку, заявление, письменная жалоба, отчет о проделанной работе, или же просто записка – любой документ, написанный от руки, предполагал грамотное написание текста. Ошибки могли повлиять на судьбу автора в разных случаях с разной степенью, но так или иначе ошибок пытались избегать.

В новом языковом пространстве – в Интернете – ситуация с правильностью использования письменного языка отличается от вышеописанной коренным образом. Поскольку Интернет предоставляет возможность не только формального общения, но и общения в высшей степени неформального, то регистры [Беликов & Крысин 2001] языка, используемые в сети, могут очень сильно варьировать. Ну а если человек общается неформально и с использованием такого регистра языка, который в принципе не предполагает существование некоторого правильного написания[1], то тем самым складываются все условия для того, чтобы освободиться от гнета правил орфографии, и единственным ограничением, которое накладывается в этом случае на процесс записи слов, остается стремление быть понятым. В самом деле, зачем писать длинное и сложное слово *сейчас* и тратить на него целых шесть букв, если можно ограничиться простым и понятным коротким *щас*, сэкономив три буквы. Конечно, если бы слово *сейчас* не редуцировалось бы до одного слова в устной речи, оно, вероятно, не могло бы быть записано таким образом. Тем не менее в этом примере мы видим как отход от произносительной нормы, так и от орфографической.

Таким образом, можно заметить, что использование письменной формы языка в таком свободном языковом пространстве, каким является Интернет, становится все больше и больше похожим на использование устного языка. Например, как видно из примера со словом *сейчас*, в этом языковом пространстве начинают действовать такие типичные закономерности устной речи, как максимы Грайса [Grice 1969]. В использовании *щас* вместо *сейчас* проявляется баланс между двумя максимами – **говори коротко** и **избегай непонятности**. Это действительно минимальное количество символов, необходимое для того, чтобы идентифицировать данное слово. Заметим, что в обычном письменном дискурсе (в котором соблюдаются в том числе правила орфографии) максимы Грайса на уровне орфографии не действуют: мы все еще вынуждены писать по правилам, например, слово *здравствуйте*, хотя такое количество букв явно излишне, как бы это слово ни редуцировалось в речи.

Итак, мы можем наблюдать, по крайней мере в таком узком языковом пространстве, как Интернет, за новым языковым процессом – за превращением одного из жанров письменной формы языка в естественный

[1] Например, такого правильного написания не предполагает редуцированная форма слова *что*. Как правильно писать: *чё* или *чо*? На этот вопрос, насколько нам известно, пока не может ответить ни один орфографический словарь.

008-Bogdanov_A:_Layout 1 18.05.2008 14:21 Page 51

язык. То есть в такую систему, которая развивается по своим внутренним естественным законам и не терпит вмешательств извне. В таком случае разумно было бы предположить, что норма в этом новом естественном языке, то есть понятие о том, что правильно, а что неправильно, может начать быстро изменяться. Так же быстро, как это происходит в обычном естественном языке, например, с фонетической нормой.

2. Описание ошибки

В этой работе мы попытались проследить за одной очень частотной орфографической ошибкой и взглянуть на нее с позиций естественности письменной формы языка и возможности становления в этой системе новой орфографической нормы.

Эта ошибка заключается в написании мягкого знака после *т* в формах единственного числа третьего лица возвратных глаголов. Несколько примеров из Интернета:

(1) а. *Мне свет солнца кажеться тусклым.*
б. *Сегодня состоиться матч Лиги Чемпионов.*
в. *Кто боиться правды?*

Эта ошибка в настоящее время действительно является очень распространенной, и особенно в текстах в Интернете. Также мягкий знак может писаться и в форме множественного числа, но мы для простоты ограничимся здесь формами единственного числа.

Отдельную интересную проблему представляет собой задача выделить тот класс глаголов (по-видимому, на основе фонологических признаков основы), для которых такая ошибка типична. Ведь можно привести примеры таких возвратных глаголов, для которых доля случаев написания с ошибкой среди всех случаев написания в Интернете много меньше соответствующей доли для приведенных в (1) глаголов. Так, например, доля написаний *понижаеться* среди всех использований этой формы в Интернете составляет 0,4 %, тогда как соответствующая доля для написания *боиться* – 3,9 %. Мы, однако, в этой работе не пытались выделить те параметры, которые могут влиять на возможность такого типа ошибок для конкретной глагольной основы.

3. Анализ частотности ошибки

Наша задача состояла в том, чтобы проанализировать динамику частотности данной ошибки. Для этого нами была использована поисковая система Яндекс. В этой поисковой системе имеется возможность ограничивать область поиска временнЫм интервалом. Собственно, анализ динамики состоял в том, что были взяты десять разных глаголов и для каждого из них произведены следующие вычисления: количество[2] употреблений данной словоформы с ошибкой в течение одного года (для всех годов с 2000 г. по 2007 г.) и количество употреблений данной словоформы без ошибки в течение одного года. Далее эти показатели суммировались и вычислялось отношение количества употреблений с ошибкой к сумме всех употреблений для каждого года. ВременнУю шкалу было решено начать с 2000 года, так как в поисковой системе Яндекс возможность поиска по дате появилась сравнительно недавно, и поэтому для более ранних периодов точность разметки ресурсов по дате уже заметно хуже, чем для периода начиная с 2000 года.

Приведем десять глагольных форм, которые были отобраны для этой работы, в написании с ошибкой:

- *боиться*
- *делаеться*
- *захочеться*
- *кажеться*
- *называеться*
- *получаеться*
- *придеться*
- *состоиться*
- *считаеться*
- *являеться*

Все эти словоформы были отобраны благодаря следующим свойствам. Ни одна из этих словоформ в написании с ошибкой не совпадает ни с одной реальной словоформой русского языка (в том числе она не совпадает и с инфинитивом своего глагола, что бывает довольно часто). Это свойство позволило осуществлять поиск по запросу состоящему только из самой этой словоформы (в кавычках), при этом была гарантия, что в

[2] В качестве показателя количества употреблений бралось количество найденных в Интернете страниц.

008-Bogdanov_A:_Layout 1 18.05.2008 14:21 Page 52

выдаче по этому запросу попадутся именно и только данные словоформы в ошибочном написании, и никакие другие словоформы языка. Также эти слова обладают сравнительно высокой частотностью, что позволяет увеличить общую точность расчета частоты.

Приведем пример одного такого расчета для одной из этих словоформ.

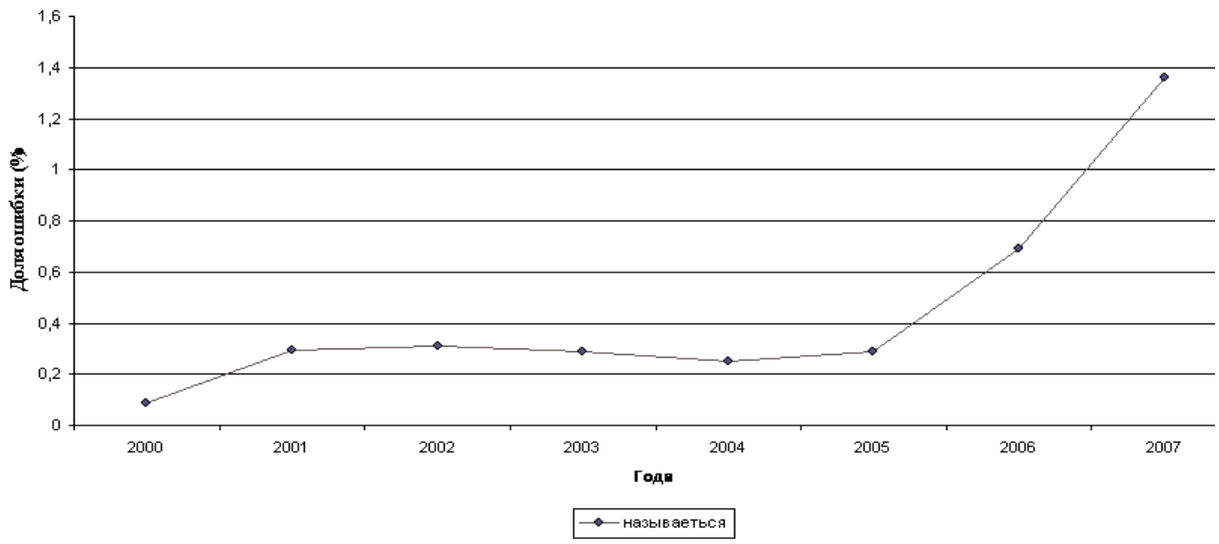
Словоформа: *называеться*

Год	Количество употреблений с ошибкой (*называеться*)	Количество употреблений без ошибки (*называется*)	Доля употреблений с ошибкой (округлено)
2000	22	25147	0,09 %
2001	121	40967	0,29 %
2002	224	71819	0,31 %
2003	410	140577	0,29 %
2004	337	135469	0,25 %
2005	629	218168	0,29 %
2006	3382	486633	0,69 %
2007	34123	2478693	1,36 %

Таблица 1.

На основе представленных в таблице 1 данных можно для наглядности изобразить такой график динамики доли ошибки для данной словоформы.

График 1 Динамика доли ошибки для словоформы *называеться*



Как видно уже только из этого графика, данные оказались весьма интересными, поскольку можно заметить, что начиная с 2005 года доля написаний с такой ошибкой начала резко возрастать.

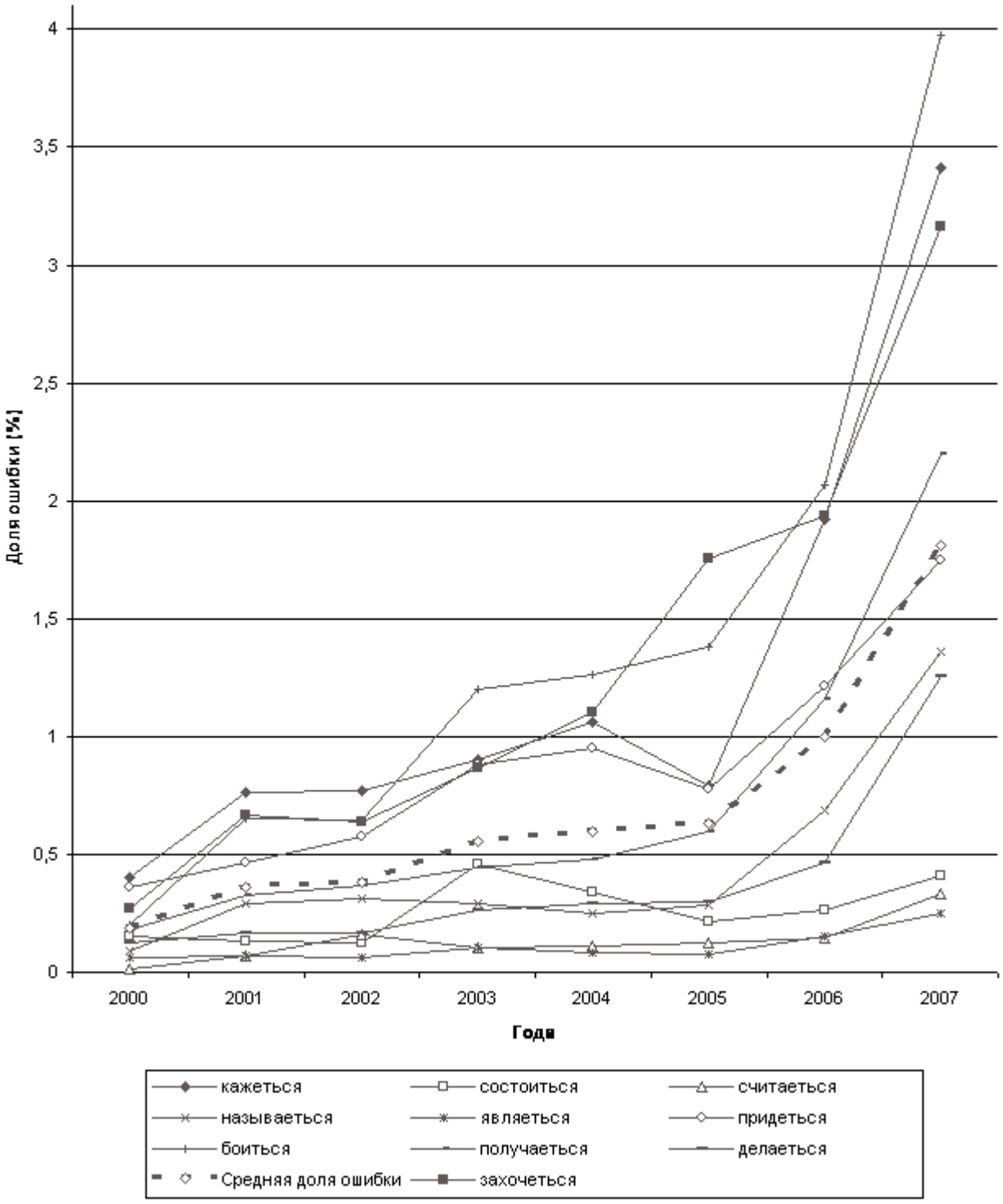
Аналогичные вычисления были проведены для каждой словоформы из списка, представленного выше. На графике 2 показана динамика доли ошибки для всех десяти словоформ из нашего списка, а также прерывистой линией изображена средняя доля ошибки.

Итак, в графике 2 показано, что по данным, вычисленным для десяти разных словоформ, была построена кривая, показывающая среднюю долю ошибки. Средняя доля считалась просто: для каждого года бралось среднее арифметическое показателей доли ошибки для каждой словоформы.

Мы, однако, не ограничились констатацией того факта, что средняя доля ошибок такого типа начиная с 2005 года начала резко возрастать. Было решено сравнить динамику для этой ошибки с общей динамикой доли орфографических ошибок.

008-Bogdanov_A:_Layout 1 18.05.2008 14:21 Page 53

График 2 Динамика доли ошибки для всех десяти словоформ



4. Анализ общей динамики орфографических ошибок

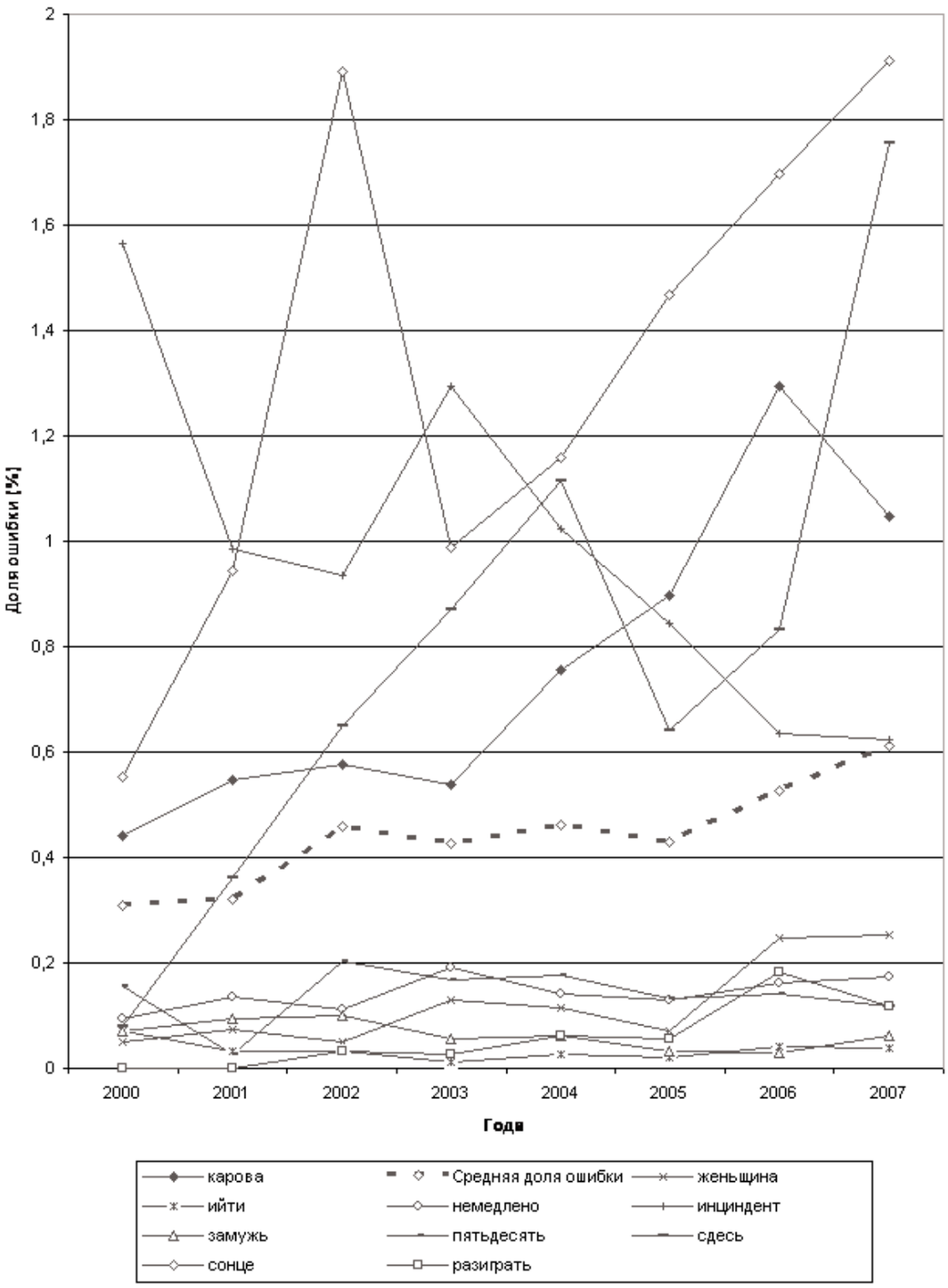
Для того, чтобы вычислить общую динамику орфографических ошибок пользователей Интернета, было взято десять типичных орфографических ошибок, не связанных с написанием мягкого знака в глагольных формах. Список состоял из следующих словоформ (приводится в записи с ошибкой).

- *карова*
- *ийти*
- *немедлено*

008-Bogdanov_A:_Layout 1 18.05.2008 14:21 Page 54

- *женьщина*
- *инциндент*
- *замужь*
- *пятьдесять*
- *сдесь*
- *сонце*
- *разиграть*

График 3 Динамика других орфографических ошибок



008-Bogdanov_A:_Layout 1 18.05.2008 14:21 Page 55

Все эти словоформы, как и глагольные словоформы из списка выше, характерны тем, что каждая из них в таком написании не является никакой реальной словоформой русского языка, что также помогло избежать шума в поисковой выдаче. К тому же они тоже являются довольно частотными словами. Также можно заметить, что все десять словоформ представляют собой примеры разных орфографических ошибок. При выборе этих словоформ мы также пытались избежать пересечения этих написаний с элементами современного сетевого жаргона (в соответствии с которым используются, например, такие написания: *превед, креведко* и т.п.). На наш взгляд, написания, представленные в списке выше, вряд ли могли бы быть основаны на использовании этого жаргона.

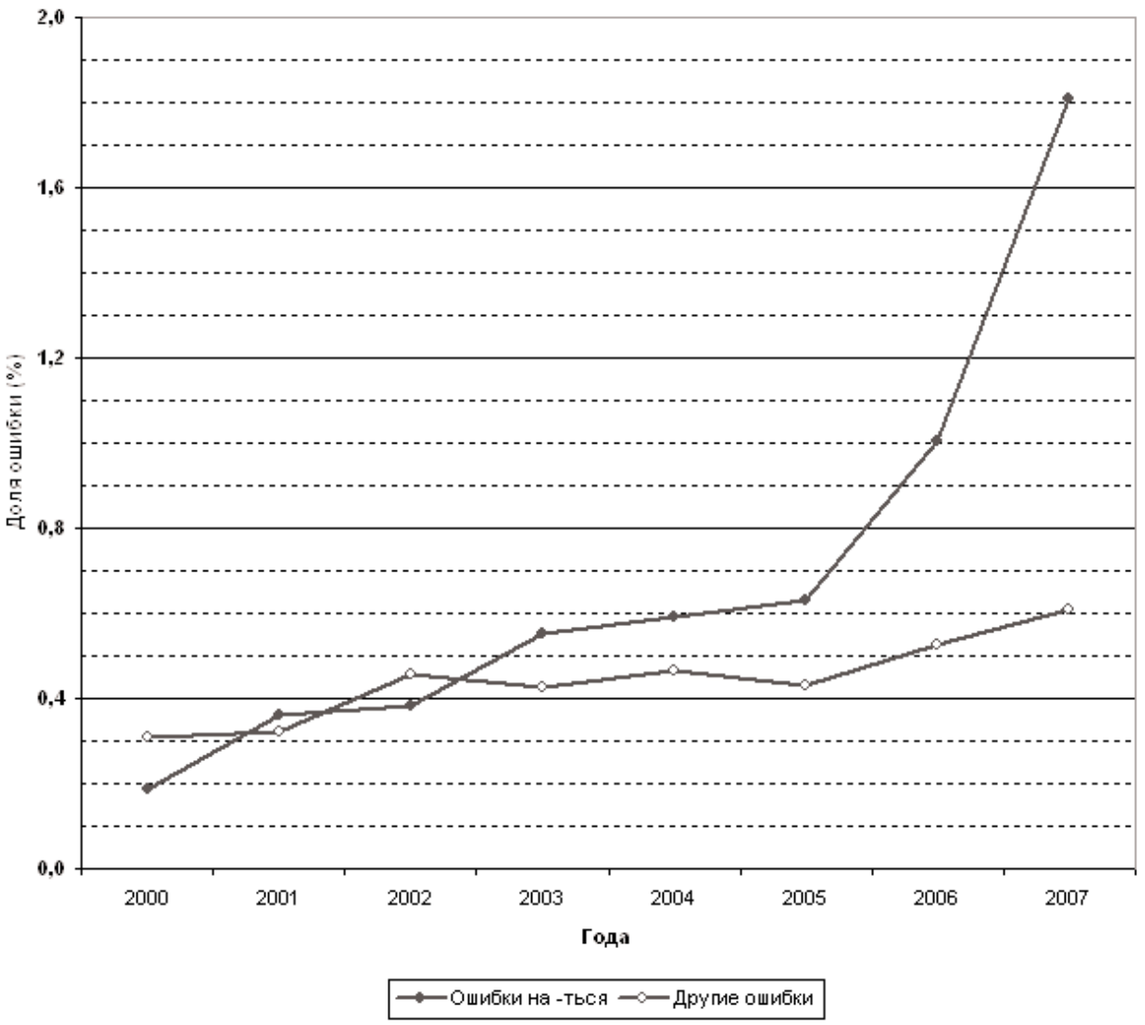
Для этих словоформ были проведены аналогичные вычисления. На графике 3 представлена динамика доли ошибки для всех десяти словоформ, а также прерывистой линией показана средняя доля ошибки.

Отметим, что при всем разнообразии поведения кривых на графике 3 значение образующих точек остается в пределах двух процентов, а среднее значение доли ошибки изменяется от 0,3 процента до 0,6 процента (округлено). Среднее значение на этом графике, как видно, тоже постепенно повышается но далеко не так резко, как среднее значение на графике 2.

5. Сравнение динамики доли ошибок на -ться с общей динамикой орфографических ошибок

Теперь нам осталось лишь сравнить две кривые, каждая из которых представляет собой усредненное значение доли ошибок. Первая – доли ошибок на *ться/тся*, вторая – доли других орфографических ошибок. На графике 4 эти две кривые представлены вместе.

График 4 Сравнение динамики доли ошибок



008-Bogdanov_A:_Layout 1 18.05.2008 14:21 Page 56

Как видно из графика 4 средние значения для доли ошибок в обоих случаях увеличиваются, но начиная с 2005 года доля ошибок на -ться начинает возрастать гораздо более резко, чем доля других ошибок. Конечно, на таком маленьком количестве данных сложно сделать достаточно достоверный вывод, но, как нам кажется, тенденцию можно рассмотреть не прибегая к более обширным исследованиям.

Итак, результатом наших вычислений стал тот факт, что доля написаний глагольных форм третьего лица единственного числа возвратных глаголов с мягким знаком не вписывается по скорости возрастания в общую тенденцию увеличения количества орфографических ошибок. На основании этих данных можно сделать вывод о некоторой иной природе таких ошибок.

Если соображения, высказанные в введении к настоящей работе, имеют право на существование и письменная форма языка в Интернете действительно начинает вести себя так же, как естественный язык, то мы вынуждены констатировать, что в данном случае мы имеем дело с новой нормой написания таких глагольных форм. В самом деле, что еще может являться доказательством становления новой нормы как не аномально возрастающая доля употребления нового варианта на фоне стабильно низкого колебания вариативности в других областях?

Конечно, данное соображение не претендует на повод для изменений в современных правилах правописания, а лишь призвано обратить внимание специалистов на природу этого явления. Однако, если тенденция, показанная нами выше, верна, то при условии возможного влияния новых норм письменного языка Интернета на «офлайновый» письменный язык, изменение действующих правил правописания может стать неизбежным.

Список литературы

1. Grice H.P. Utterer’s meaning and intentions // Philosophical Review, 1969. Vol.78. P. 54 – 70
2. Беликов В.И. & Крысин Л.П. Социолингвистика, М., 2001

009-Bogdanova1:_Layout 1 18.05.2008 14:21 Page 57

О «КОРПУСЕ» ТЕКСТОВ ЖИВОЙ РЕЧИ: ПРИНЦИПЫ ФОРМИРОВАНИЯ И ВОЗМОЖНОСТИ ОПИСАНИЯ

THE CORPUS OF SPOKEN RUSSIAN: DESIGN PRINCIPLES AND APPROACHES TO DATA ANALYSIS

Богданова Н.В. *(nvbogdanova_2005@mail.ru),* ***Бродт И.С.*** *(brodt_05@mail.ru),* ***Куканова В.В.*** *(vika.kukanova@gmail.com),* ***Павлова О.В.*** *(olgapavlovaspb@mail.ru),* ***Сапунова Е.М.*** *(kaverita@yandex.ru),* ***Филиппова Н.С.*** *(ninaphilippova@gmail.com)*
Санкт-Петербургский государственный университет

В докладе рассматриваются принципы формирования своеобразного звукового корпуса – массива текстов живой русской речи, объединенных едиными лингвистическими и социолингвистическими параметрами. Описываются готовые блоки такого корпуса, возможности его многоуровневого описания и перспективы расширения.

Одним из актуальных и активно развивающихся направлений современной лингвистики является сбор и систематизация живого речевого материала[1]. Этим занимается *полевая лингвистика*, под которой понимают «лингвистическую дисциплину, разрабатывающую и практикующую методы получения информации о неизвестном исследователю языке на основании работы с его носителями» (Кибрик 2007). В настоящем исследовании методы полевой лингвистики используются для построения и описания своеобразного «корпуса» текстов живой русской речи, о которой, как показывают первые исследования на этом материале, лингвистике известно примерно так же мало, как о каком-нибудь действительно неизвестном языке, и любая попытка описать эту речь с применением тех методов и того лингвистического инструментария, который традиционно используется для анализа письменно-литературного языка, наталкивается на сопротивление самого материала, что вынуждает исследователей ставить и решать массу абсолютно новых задач, начиная с определения самóго метаязыка такого лингвистического описания (см. *Полевая лингвистическая практика 2008*). Иными словами, к спонтанной речи – с большей или меньшей степенью уверенности – мы можем относиться, как к новому объекту изучения, и применять к нему устоявшиеся методы полевой лингвистики. А. Кибрик противопоставляет полевую лингвистику «кабинетной», на том основании, что источником данных для последней «являются либо языковая интуиция самого исследователя, являющегося носителем изучаемого языка или, по крайней мере, хорошо им владеющего, либо обширный корпус текстов на изучаемом языке, о котором опять же известно достаточно много для того, чтобы изучать его без обращения к суждениям его носителей» (*Кибрик 2007*). В случае с живой речью, с одной стороны, исследователи сами являются ее носителями, и даже «хорошо ею владеющими», но, с другой, не могут изучать ее без обращения к другим носителям. Именно эти неискушенные носители языка – информанты – являются «посредниками между исследователем и языком», а задача исследователя – «эффективно воздействовать на языковую деятельность информанта», чтобы получить от него тот или иной речевой продукт. Для получения такого продукта полевая лингвистика использует «активный метод целенаправленного интервьюирования по определенной программе», в ходе осуществления которого языковая деятельность информанта протекает максимально естественно и спонтанно.

Именно такая программа легла в основу формирования звукового «корпуса» спонтанных текстов на русском языке, создаваемого нашим научным коллективом. Слово «корпус» употреблено в данном случае до некоторой степени условно, речь идет скорее о своеобразной текстотеке, о массиве живых текстов на русском языке, у которых есть ряд общих и ряд различных черт и которые могут быть, с одной стороны, материалом для самых разных исследований (от фонетики до функциональной стилистики, в том числе в социо- и психолингвистическом аспектах), а с другой – частью того самого корпуса живой русской речи, создание которого и является целью современной полевой лингвистики.

[1] См. об этом: *Полевая лингвистическая практика 2007*. В качестве примера таких баз данных можно назвать, кроме того, проект «Отчеты детей об их сновидениях» (руководители А. А. Кибрик и В. И. Подлесская)», а также устную часть Национального корпуса русского языка.

009-Bogdanova1:_Layout 1 18.05.2008 14:21 Page 58

С точки зрения функциональной принадлежности можно говорить о монологах различного характера. Чаще всего исследователи обращают свой взор на устную публичную речь – общественно-политическую, деловую или научную. В противоположность этому нас интересует только *бытовой спонтанный монолог*, характеризующийся неподготовленностью, непринужденностью, неофициальностью и необязательным участием говорящего в акте коммуникации (роль второго участника такого акта сводится к минимуму, хотя, безусловно, никогда не исчезает полностью).

Непременное требование спонтанности, предъявляемое к материалу фиксации и исследования, не отменяет того факта, что степень такой спонтанности может быть различной и зависит от многих факторов.

Главным из таких факторов следует назвать степень *лингвистической мотивированности* речевого произведения, то есть обусловленности (тематической и лингвистической) монолога как продукта речевой деятельности человека (вторичного текста) характеристиками некоторого исходного, первичного, ставшего стимулом для появления вторичного. Разнообразие таких стимулов велико: от вопроса, ответом на который становится развернутый спонтанный монолог (минимальная степень мотивированности и максимальная степень спонтанности), до другого текста (предтекста), предназначенного для прочтения (максимальная степень мотивированности и минимальная степень спонтанности) или пересказа (более низкая, чем при прочтении, степень лингвистической мотивированности и, соответственно, более высокая степень спонтанности). Промежуточное положение на этой шкале занимает описание зрительного ряда (изображения), обладающее средней степенью мотивированности и такой же средней степенью спонтанности.

Лингвистическая мотивированность и спонтанность речевого произведения (в нашем случае – бытового монологического текста) находятся в отношениях обратно пропорциональной зависимости: увеличение степени мотивированности ведет к снижению спонтанности, и наоборот[2].

Дополнительными характеристиками исходного стимула, способными повлиять на свойства спонтанного монолога, стали в нашем исследовании сюжетность/несюжетность предтекста или изображения или степень знакомства говорящего с темой свободного монолога, заданного вопросом (исходным стимулом). Эти дополнительные характеристики не меняют степени лингвистической мотивированности и, соответственно, спонтанности вторичного текста, но все же оказывают влияние на выбор говорящим речевых средств и в целом на лингвистическую природу вторичного текста. Можно предположить, что в данном случае решающими являются характеристики уже не (или не только) первичного текста, но и самого говорящего – уровень его речевой компетенции (УРК) или психологический тип его личности.

Наш корпус, таким образом, составляют спонтанные бытовые монологи разной степени лингвистической мотивированности и спонтанности, записанные от носителей русского языка с различными социальными характеристиками. Последние представлены главным образом такими показателями:

- пол,
- возраст,
- профессиональная принадлежность,
- профессиональное или непрофессиональное отношение к речи,
- уровень речевой компетенции.

Думается, что требуют комментариев два последних признака.

Профессиональное или непрофессиональное отношение говорящего к речи устанавливается через определение роли речи (языка) в его жизни. Здесь возможно несколько вариантов:

1) речь для человека – только *средство коммуникации* (большинство носителей языка – не школьники, не филологи, не преподаватели, не актеры, не лекторы и т. п.);

2) речь – *средство коммуникации и объект изучения* (до некоторой степени школьники, а также студенты-филологи и «кабинетные» ученые-филологи);

3) речь – *средство коммуникации* и *орудие труда* (преподаватели-нефилологи, актеры, лекторы, публичные и общественные деятели, политики и т. п.)[3];

4) речь – *средство коммуникации, объект изучения* и *орудие труда* (преподаватели-филологи; особое место среди них, как представляется, занимают преподаватели русского языка как иностранного).

Уровень речевой компетенции говорящего определяется целой совокупностью его социальных характеристик, среди которых ведущее место занимают уровень образования, профессиональное или непрофессиональное отношение к речи, а также степень социальной активности личности. Этот набор признаков, определяющих УРК, был в свое время установлен экспериментальным путем – через экспертную

[2] См. подробнее об этой типологии Богданова 2004; Полевая лингвистическая практика 2008.

[3] Число таких носителей языка в современном мире неуклонно увеличивается, что ставит перед специалистами отдельную задачу обучения устной публичной речи. Этой цели может служить, в частности, и создание различных корпусов (текстотек) живой речи, записанных, в частности, от носителей языка с высоким УРК.

009-Bogdanova1:_Layout 1 18.05.2008 14:21 Page 59

оценку большого массива звучащих текстов и выявление корреляции этих оценок с реальными социальными характеристиками говорящих[4]. В нашем случае мы просто опирались на эти признаки и на их основе определяли УРК наших информантов, хотя проведение экспертной оценки и уточнение проведенного разделения на группы вполне возможно.

Общую характеристику нашего материала удобно дать, отталкиваясь от определения *национального корпуса*, принятого в современной корпусной лингвистике, – это «информационно-справочная система, основанная на собрании текстов в электронной форме. Национальный корпус представляет данный язык на определенном этапе (или этапах) его существования и во всем многообразии жанров, стилей, территориальных и социальных вариантов и т. п. <…> Национальный корпус создается лингвистами (специалистами по так называемой корпусной лингвистике, быстро развивающейся современной области языкознания) для научных исследований и обучения языку. <…> Национальный корпус имеет две важные особенности. Во-первых, он характеризуется представительностью, или сбалансированным составом текстов. Это означает, что корпус содержит по возможности все типы письменных и устных текстов, представленных в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т. п.), и что все эти тексты входят в корпус по возможности пропорционально их доле в языке соответствующего периода. Во-вторых, корпус содержит особую дополнительную информацию о свойствах входящих в него текстов (так называемую разметку, или аннотацию). Разметка – главная характеристика корпуса; она отличает корпус от простых коллекций (или “библиотек”) текстов. Чем богаче и разнообразнее разметка, тем выше научная и учебная ценность корпуса»[5]. Приложим конкретные характеристики национального корпуса к нашему материалу.

1. *«Информационно-справочная система, основанная на собрании текстов в электронной форме»*. Наш материал представляет собой некоторую основу для настоящей *«информационно-справочной системы»*, – это ***собрание текстов*** в виде магнитных записей и их орфографических расшифровок. Часть материала существует уже и в электронном виде, возможность перевода в эту форму остальных текстов, разумеется, существует.

2. Корпус представляет *«...язык во всем многообразии жанров, стилей, территориальных и социальных вариантов и т. п.…»*. Выше уже была представлена наша ***типология текстов***, не претендующая, безусловно, на всеохватность, но соответствующая тем теоретическим установкам, на которых строится анализ собираемого нами материала.

3. Корпус характеризуется *«представительностью, или сбалансированным составом текстов»*. В рамках выдвинутой гипотезы о существовании корреляции между лингвистическими характеристиками спонтанного монолога и его типом, с одной стороны (собственно лингвистический аспект исследования), а также между этими характеристиками и социальными и психологическими признаками говорящих, с другой (психо- и социолингвистический аспекты исследования), вполне можно, думается, говорить о ***представительности*** и ***сбалансированности***.

4. Корпус *«содержит особую дополнительную информацию о свойствах входящих в него текстов (так называемую разметку, или аннотацию)»*. В нашем случае такой дополнительной информацией (и одновременно разметкой) является прежде всего *синтаксическое пунктирование текстов*, полученное для части материала в ходе специальных экспериментов с привлечением экспертов-филологов. Подобное пунктирование дает в руки исследователю некую единицу описания, соотносимую с традиционным предложением и позволяющую осуществить дальнейший синтаксический анализ спонтанных монологов во всей его полноте[6].

Другим вариантом синтаксической разметки звучащего материала стало выделение в спонтанном тексте *стуктурно-синтаксических единств* (ССЕ), под которыми понимаются связанные комплексы – «предикативные центры (полнозначные слова в функции главных членов предложения и нечленимые слова-предложения), сами по себе или с зависимыми словоформами и полупредикативными конструкциями» (*Филиппова 2006*).

Другим типом разметки на нашем материале является отражение в расшифровках того, что в определении корпуса обозначено как *«изменение качества речи, паузация и разнообразные паралингвистические явления (например, смех) в устной речи»*. Все это (а также повторы, обрывы речи, самоперебивы и самокоррекция, паузы хезитации – как неотъемлемые признаки любого спонтанного монолога) присутствует в зафиксированном материале, а в некоторых случаях даже стало объектом специального рассмотрения.

Наличие в нашем материале подобной разметки уже позволяет говорить о нем, как о некоем подобии корпуса, поскольку именно *«она отличает корпус от простых коллекций (или “библиотек”) текстов»*.

5. Наш материал представляет собой своеобразное объединение того, что в определении корпуса называется *«демографической частью»* (спонтанная речь повседневного общения) и *«контекстно-*

[4] См. в списке использованной литературы серию отчетов ЛЭФ им. Л. В. Щербы за 1985-90 гг.
[5] Что такое Корпус? // Официальный сайт Национального корпуса русского языка. http://www.ruscorpora.ru/corpora-intro.html.
[6] См. об этом подробнее *Богданова 2006, Бродт 2007*.

009-Bogdanova1:_Layout 1 18.05.2008 14:21 Page 60

ориентированной» устной речью, – см. предложенную выше типологию составляющих его спонтанных монологов.

6. Наш материал – это, безусловно, *«корпус общего типа»*, т. к. он содержит разные речевые жанры.

В целом возможные *аспекты использования* собранного материала можно представить следующим образом:

1) собственно *лингвистические исследования*:

• специфика устной спонтанной речи на всех уровнях;

• пересмотр нормативных требований к построению живого (в первую очередь устного) монологического текста;

• создание лексикографического описания бытовой спонтанной звучащей речи;

• описание дистрибуции тех или иных грамматических классов слов или их форм в устной монологической речи разных социальных групп;

2) *лингводидактика*:

• обучение русскому языку нерусских; собрание дает богатый материал для учебного аудирования и вообще знакомства с живой речью в иностранной аудитории;

• изучение грамматики речи в русской филологической аудитории;

3) материал для *психо- и социолингвистических исследований*;

4) материал для исследований в области *коллоквиалистики*;

5) *прикладная лингвистика*:

• решение задач обработки естественного языка/речи;

• решение задач интегрального моделирования звуковой формы.

Основная единица описания в собрании – *спонтанный монологический текст* в звучащем и расшифрованном виде.

Состав информантов – носителей русского языка, от которых записан весь наш материал, – будучи весьма разнородным по социальным и психологическим характеристикам говорящих (это было одним из непременных условий всех проведенных экспериментов), является, тем не менее, строго однородным в территориальном отношении: все информанты являются коренными петербуржцами, т. е. носителями петербургского произносительного варианта современного русского литературного языка.

Объем и состав «корпуса»:

1) спонтанные устные монологи разного типа (здесь и далее речь идет о предложенной выше типологии), записанные от 30 информантов, женщин-медиков одной возрастной группы, но с разным УРК (210 текстов, около 9 час. звучания) (*Бродт 2007*);

2) спонтанные устные монологи разного типа, записанные от 6 информантов, преподавателей русского языка как иностранного (наиболее высокий УРК), разного пола и возраста (42 текста, около 2 час. звучания) (*Павлова 2007*);

3) спонтанные устные монологи разного типа, записанные от 43 информантов, мужчин-юристов разного возраста и разного УРК (301 текст, около 12 час. звучания) (*Куканова 2007; 2008*);

4) спонтанные устные монологи – описания сюжетного и несюжетного изображения, – записанные от 20 информантов одного возраста (19 22 года) и приблизительно одного УРК (студенты – филологи и нефилологи) (40 текстов, около 2 час. звучания) (*Филиппова 2006; 2007*);

5) спонтанные устные монологи – свободные рассказы на заданную тему (знакомую и незнакомую), – записанные от 20 информантов одного возраста и приблизительного одного УРК (студенты – филологи и нефилологи) (40 текстов, около 2 час. звучания) (*Конюхова 2006*);

6) неподготовленное (спонтанное) чтение двух текстов (сюжетного и несюжетного), записанное от 12 информантов разного пола, но одной возрастной группы и одного – среднего – УРК (студенты – филологи и нефилологи) (24 текста; 1 час звучания) (*Сапунова 2007)*;

Как видно, общий объем материала, ставшего частью создаваемого массива («корпуса») живых текстов на русском языке, достаточно представителен и разнообразен. Столь же разнообразны и возможности его расширения и практического использования.

009-Bogdanova1:_Layout 1 18.05.2008 14:21 Page 61

Список литературы

1. Богданова Н.В. Типология спонтанных монологов в устной и письменной формах речи // Фонетические чтения. К 100 летию Л. Р. Зиндера. СПб.: 2004. С. 214 217.

2. Богданова Н.В. О единице описания синтаксической структуры устного спонтанного монолога: проблемы, методики, гипотезы // …СЛОВО ОТЗОВЕТСЯ. Памяти А. С. Штерн и Л. В. Сахарного. Пермь: 2006. С. 288 293.

3. Бродт И.С. Спонтанный монолог в лингвистическом и социолингвистическом аспектах (на материале текстов разного типа). Дис. … канд. фил. наук, СПб.: 2007.

4. Исследование отражения в речи некоторых социальных характеристик говорящего. Отчеты ЛЭФ. Л., 1987, 1988, 1990.

5. Кибрик А. Полевая лингвистика // www.krugosvet.ru/articles/77/1007704/1007704a1.htm (2007).

6. Конюхова А.А. Лингвистические особенности свободного монолога на заданную тему // Русская филология. 18. Сборник научных работ молодых филологов. Тарту: 2007. С. 222 230.

7. Куканова В.В. Об одном из способов подбора информантов в ходе полевого исследования // Материалы XXXVI международной филологической конференции. Выпуск 20. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 12 17 марта 2007 года, СПб.: 2007. С. 45 52.

8. Куканова В.В. Русская спонтанная речь. Методическая разработка по современному русскому языку. Выпуск I. Свободные монологи-рассказы на заданную тему. Тексты. СПб.: 2008 (в печати).

9. Методика получения языкового материала для изучения социальной характеристики говорящего. Отчет ЛЭФ. Л., 1985.

10. Павлова О.В. Влияние возраста говорящего на синтаксические характеристики его речи // Материалы XXXVI международной филологической конференции. Выпуск 20. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 12 17 марта 2007 года, СПб.: 2007. С. 52 59.

11. Полевая лингвистическая практика. Учебно-методический комплекс сложной структуры. Часть 1. Теоретические основы и методика сбора лингвистических данных для представления их в речевом корпусе русского языка / Ред. Асиновский А. С., Богданова Н. В. СПб., 2007.

12. Полевая лингвистическая практика. Учебно-методический комплекс сложной структуры. Часть 2. Методические указания по обработке, многоуровневой разметке и лингвистическому анализу корпуса звучащих текстов на русском языке / Ред. Асиновский А. С., Богданова Н. В. СПб.: 2008 (в печати).

13. Сапунова Е.М. Неподготовленное чтение как разновидность устного спонтанного монолога // Материалы XXXVI международной филологической конференции. Выпуск 20. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 12 17 марта 2007 года, СПб.: 2007. С. 76 86.

14. Филиппова Н.С. Опыт анализа синтаксической структуры устного спонтанного монолога-описания // IX Межвузовская научная конференция студентов-филологов. Тезисы. 10 14 апреля 2006 г. Санкт-Петербург. СПб.: 2006. С. 51 52.

15. Филиппова Н.С. Операции отмены как способ организации спонтанной речи (на материале устных спонтанных монологов-описаний) // Материалы XXXVI международной филологической конференции. Выпуск 20. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 12 17 марта 2007 года, СПб.: 2007. С. 86 90.

16. Что такое Корпус? // Официальный сайт Национального корпуса русского языка. http://www.ruscorpora.ru/corpora-intro.html.

010-Borschew_V:_Layout 1 18.05.2008 14:21 Page 62

«Я НЕ БЫЛ..., МЕНЯ НЕ БЫЛО...», ИЛИ СКОЛЬКО РАЗНЫХ *БЫТЬ* В РУССКОМ ЯЗЫКЕ*

"JA NE BYL…MENJA NE BYLO.." OR HOW MANY DIFFERENT *BYT'* (*BE*) IN RUSSIAN

Борщев В.Б. *(borschev@linguist.umass.edu)*
ВИНИТИ РАН; UMass

В работе приводится и анализируется пример Я не был в зале, когда выключили свет. Пример этот опровергает утверждение Ю.Д. Апресяна, что для такого рода предложений невозможна «синхронная» интерпретация. Обсуждаются разные значения глагола быть в локативных и бытийных предложениях.

1. Введение

1.1. Тема

Есть известный пример Апресяна, относящийся к локативному значению глагола *быть* (Апресян 1980, примеры (107i –107iii):

(1) *Отец был на море*.
(2) *Отец не был на море*.
(3) *Отца не было на море*.

Апресян пишет[1]: «Фразы типа (1) омонимичны: они могут значить либо то, что отец в принципе бывал на море (общефактическое нерезультативное значение НЕСОВ), либо то, что отец в какой-то фиксированный момент находился на берегу моря (процессное значение НЕСОВ).

Под отрицанием возможны две разных формы таких фраз – с сохранением именительного падежа подлежащего (2) и с изменением именительного падежа на родительный (3).

В отличие (1), фразы (2) и особенно (3) неомонимичны...».

Заметим, что Апресян с некоторой осторожностью пишет о том, что фраза (2) неомонимична.

Этот пример Апресяна (предложения (1) - (3)) существен для описания семантики глагола *быть*. Не случайно, что он повторяет его в двух других своих работах (Апресян 1992 и Апресян 2005)

Когда мы с Барбарой Парти начали заниматься генитивом отрицания (Gen Neg), нам казалось, что глагол *быть* в прошедшем времени может иметь значение, которое Апресян называет «процессным значением НЕСОВ»[2] – надо только найти подходящий контекст, как-то «фиксировать момент» бытия... Приходили в голову примеры типа (4):

(4) *Петя не был тогда/в тот момент в Москве.*

Однако большинство наших коллег относилось к такого рода примерам скептически, утверждая, что и тут возможна «динамическая» трактовка: не был – значит не побывал. «Чистый» пример долго не удавалось найти.

1.2. Мой пример

В 2005 г. мне удалось найти такой пример:

(5) *Я не был в зале, когда выключили свет*.

Действительно, предложение (5) фиксирует момент времени, исключая тем самым значение ‘в принципе бывал’, которое Апресян сопоставляет фразе (2).

* Работа была частично поддержана грантом NSF (National Science Foundation Grant No. BCS-0418311 to Partee and Borschev, "The Russian Genitive of Negation: Integration of Lexical and Compositional Semantics").

[1] В приводимой цитате, как и последующих, изменена нумерация примеров – она согласована с нумерацией примеров в данной статье.

[2] В работах (Апресян 1992 и Апресян 2005) это значение называется «актуально длительным».

010-Borschew_V:_Layout 1 18.05.2008 14:21 Page 63

Для предложения (5) возможен и его «генитивный» вариант (6), т.е. конструкция с Gen Neg, а также аналогичные предложения (7) и (8) без личных местоимений[3]:

(6) *Меня не было в зале, когда выключили свет*.
(7) *Петя не был в зале, когда выключили свет*.
(8) *Пети не было в зале, когда выключили свет*.

Естественно считать, что в предложениях (5) – (8) глагол *быть* употребляется в одном и том же значении. Собственно, цель этих примеров – понять, в каких значениях глагол *быть* употребляется в конструкции Gen Neg и, шире – в бытийных и локативных предложениях.

1.3. Падучева и мой пример

Я благодарен Е.В. Падучевой за популяризацию моего примера. Она привела его в нескольких статьях (см., например, Падучева 2006a, 2006b, 2008). Кроме того, в двух других статьях (Падучева, рукопись 2007, Paducheva, in print) она привела свои примеры такого рода предложений, см. пример (9) ниже:

(9) *Я, к счастью, не был в Грозном/Махачкале, когда там началась эпидемия холеры.*

К сожалению, я не всегда согласен с ее анализом. Цель этой заметки – привести анализ этих предложений, который представляется мне правильным, и обсудить значения, которые принимает в них глагол *быть*.

2. Локативное быть (а также экзистенциональное и посессивное)

2.1. Некоторые существующие подходы

Я кратко, скорее даже пунктирно остановлюсь только на двух отечественных.

Ю.Д. Апресян (Апресян 1980 и, прежде всего, Апресян 1992) выделяет у глагола *быть* «шесть крупных групп значений: ...2. локативные, 3.посессивные, 4. экзистенциональные ...», каждая из которых, в свою очередь, «представлена несколькими значениями...».

Примеры двух из них:

«2.1. *А был в <на> Х-е* = '*А* находился в <на> *Х-е* ...

2.2. *А был в <на> Х-е* = 'Человек *А* переместился в <на> *Х* и находился там для занятий какой-л. деятельностью' [… в формах ПРОШ имеет общефактическое двунаправленное значение: 'пришел <приехал> в *Х*, а потом ушел <уехал > из него'; СИН: *посещать*...]».

Анализ Апресяном предложений (1) - (3) является иллюстрацией этого подхода.

Е.В. Падучева говорит о двух локативных значениях глагола *быть*, двух лексемах — *быть*$_1$ и *быть*$_2$. Первую она называет *статической*, а вторую *динамической*. Про вторую она пишет: «... *быть*$_2$, как в (3), значит, приблизительно, 'побывать', т.е. переместиться в Место, находиться там в течение некоторого времени и вернуться.».

Можно отметить сходство этих двух подходов: оба выделяют по крайней мере две разные лексемы, «статическую» и «динамическую (двунаправленную)».

2.2. Оговорки и пожелания

Я не лексикограф, более того, даже не лингвист. И я понимаю всю сложность такого рода работы, особенно относящейся к такому особому глаголу, как *быть*. Поэтому все мои пожелания следует пометить столь популярным теперь ярлыком ИМХО (very humble…).

Все же, учитывая оговорки. Мне не хотелось бы разделять не только локативные значения *быть* на отдельные лексемы, но и отделять их от экзистенциональных и посессивных значений. А для описания значений соответствущих предложений стараться использовать, по возможности, универсальные, системные средства (в противовес, или, скорее, в дополнение к словарным).

2.3. Несколько конкретных примеров

Мне не хотелось бы разделять «статическое» и «динамическое» значения локативного *быть* (типа *быть*$_1$ и *быть*$_2$ у Падучевой или 2.1 и 2.2 у Апресяна), я не вижу лексической разницы между ними, не вижу

[3] После появления примера (1) мы с Е.В. Падучевой активно обсуждали близкие примеры, пытаясь нащупать границы допустимого. Я уже не помню, кто предложил примеры типа (6) – (8).

010-Borschew_V:_Layout 1 18.05.2008 14:21 Page 64

перемещения и возвращения или «двунаправленного значения», синонимичного *посещать*. Хотя, конечно, я согласен с тем, что значения предложений (2) и (3) демонстрируют соответствующую разницу. Но разницу в значении этих предложений можно, как мне кажется, объяснить другими, более универсальными средствами.

При этом можно опереться на работы Апресяна и Падучевой. Так Падучева в работах (Падучева 2006а. и 2008), приводит примеры (10) и (11) из работы (Апресян 1980), давая им несколько другую интерпретацию:

(10) *Вот на этой стене висела картина*. <Где она сейчас?>
(11) <Коля оглядел комнату.> *На стене справа висела картина*

В этих примерах по-разному интерпретируется прошедшее время. В терминологии Рейхенбаха в предложении (10) время события (event time) предшествует точке отсчета (reference time), картина висела когда-то и не висит в тот момент, когда смотрят на стену. А для предложения (11) эти моменты совпадают. Заметим, что никто не считает, что в этих предложениях используются разные лексемы глагола *висеть*.

Но ведь омонимия предложения (1) в примере Апресяна связана с той же разницей в интерпретации прошедшего времени. Зачем же здесь вводить разные лексемы глагола *быть*?

Точно также «динамика», «двунаправленность» и т.п. легко объясняются универсальными принципами, в данном случае аксиомами Грайса. По этим аксиомам в предложении типа (10) достаточно сказать, что картина *висела*, а то, что она не висит, когда смотрят на стену – это импликатура. Точно также при ретроспективной интерпретации предложения (1) не нужно говорить о том, что отец сейчас не на море, достаточно сказать, что он там *был*. «Динамика», «двунаправленность» – это универсальные импликатуры при рассмотренной выше интерпретации прошедшего времени, а не особое лексическое значение глагола *быть*. Повторяю, что Падучева сама писала об этом.

2.4. Структура перспективы

Для дальнейшего мне понадобится понятие структуры перспективы, которое мы ввели в наших работах с Барбарой Парти, занимаясь семантикой бытийных и локативных предложений.

Мы считали, что для различения ситуаций, описываемых такими предложениями, недостаточно учитывать их синтаксическую структуру и структуру Темы-Ремы. Поэтому в нескольких наших работах, в том числе в уже цитированной статье (Борщев и Парти 2002) мы ввели понятие структуры перспективы для различения такого рода ситуаций.

Прежде всего, следуя Арутюновой и Ширяеву, мы выделяли в структуре бытийных (и локативных) предложений три компонента: (бытийный) глагол и два его аргумента, которые мы называли ВЕЩЬ и МЕСТО. Т.е. по сути дела мы схематизировали ситуации бытия, описываемые бытийными и локативными предложениями, выделяя в них двух участников, ВЕЩЬ и МЕСТО, в котором это бытие происходит, и «отношение бытия» между этими участниками ситуации. Схема эта аналогична предложению Jackendoff'a (1972) о метафорическом расширении "being in some location", включающем в себя "being in some state", "occurring in some spatiotemporal region", "being in someone's possession" (а также бытие в «перцептивном пространстве Субъекта сознания» – Падучева 1997).

При этом мы подчеркивали роль МЕСТА в классификации бытийных и локативных предложений.

Так в работе (Борщев и Парти 2002) мы писали: "МЕСТО может быть указано явно или подразумеваться... Важно, что бытие всегда мыслится в некотором МЕСТЕ. Подразумеваемое МЕСТО задается контекстом. Это обычно «здесь» или «там», «сейчас» или «тогда»... Иногда ... МЕСТО естественно отождествлять «со всем миром». Итак, мы считаем, что как бытийные предложения с глаголом *быть*, так же, как и их декларативные аналоги, описывают ситуацию бытия, утверждая или отрицая бытие ВЕЩИ в МЕСТЕ" (Борщев и Парти 2002:68).

Мы считали, что локативные и бытийные предложения различаются *выбором перспективы*, точки зрения, с которой мы рассматриваем соответствующие ситуации. Выбор перспективы – это, так сказать, разное когнитивное представление той или иной ситуации. Мы можем описывать ситуацию бытия в локативном предложении с точки зрения ВЕЩИ, а в бытийном предложении с точки зрения МЕСТА. В первом случае мы как бы следим за ВЕЩЬЮ, предполагаем ее существование и говорим о ней, о ее, так сказать, поведении в ситуации, о связанных с ней действиях или о ее состоянии. Во втором случае мы смотрим на ситуацию с точки зрения МЕСТА и следим в первую очередь за тем, какие ВЕЩИ есть, а каких нет в этом МЕСТЕ. Т.е. МЕСТО является, в каком-то смысле «семантическим центром» таких предложений, в них как бы предицируются свойства МЕСТА.

Мы пользовались метафорой (кино)камеры. Обозревая ситуацию, камера может быть сфокусирована на «герое» (ВЕЩИ в нашей терминологии), следить за ним и его жизнью, в частности, за тем, в каком МЕСТЕ он

010-Borschew_V:_Layout 1 18.05.2008 14:21 Page 65

находится или отсутствует, а может быть сфокусирована на этом МЕСТЕ и наблюдать, какие герои появляются (или не появляются) в нем.

Формально мы сводили понятие перспективы к выделению (маркированию) одного из этих участников ситуации, участника, с точки зрения которого мы эту ситуацию рассматриваем. Мы называли выделенную роль *центром перспективы*.

Выбор перспективы в ситуации определяет, какое предложение – локативное или бытийное – будет выбрано для описания этой ситуации. Выделяя ВЕЩЬ (выбирая ее центром перспективы), мы описываем ситуацию локативным предложением, а выделяя МЕСТО – бытийным.

Форма отрицательных бытийных предложений – генитив субъекта и безличное сказуемое – определяется выбором перспективы (тем, что центром перспективы является МЕСТО).

С точки зрения структуры дискурса, локативное предложение (с глаголом *быть*) имеет смысл рассматривать, как сообщение о том, что ВЕЩЬ, именуемая подлежащим, имеет свойство «быть (или не быть) в данном МЕСТЕ», т.е предикат бытия в данном МЕСТЕ выполняется или не выполняется для ВЕЩИ. А бытийное предложение, это сообщение о том, что МЕСТО обладает свойством, что там есть ВЕЩЬ (нет ВЕЩИ), т.е предикат существования ВЕЩИ выполняется или не выполняется для данного МЕСТА.

Замечание. Нужно сказать, что далеко не всегда одну и ту же ситуацию можно представить двумя способами, маркируя одного из двух участников. В работах по Gen Neg часто рассматриваются пары предложений, бытийное и локативное, например, предложение (12), и его локативный аналог, предложение (13):

(12) *Посуды на столе не стояло*.
(13) *Посуда на столе не стояла*.

Эти предложения описывают сходные, но все же разные ситуации. В предложении (13) с референтным субъектом речь идет о конкретной посуде (которая не стояла на столе, а была где-то), а в предложении (12) говорится просто о том, что на столе не было посуды. Описывя разные ситуации, предложения эти обладают и разной структурой перспективы, в предложении (12) маркировано МЕСТО, а в предложении (13) маркирована ВЕЩЬ.

2.5. Предложения (5) – (8) и структура перспективы

В предложениях (5) и (7) центром перспективы является ВЕЩЬ, а в предложениях (6) и (8) – МЕСТО. В работе (Борщев, Парти. В Печати) мы обсуждали, как естественно классифицировать предложения (6) и (8). В каком-то смысле они обладают свойствами как бытийных, так и локативных предложений.

Обсуждая границы примера (5), Падучева в работе (2008) рассматривает примеры (14) и (15). Она при этом опирается на понятие Наблюдателя, в чем-то аналогичное понятию структуры перспективы:

(14) ?*Я не был дома, когда выключили свет*
(15) *Меня не было дома, когда выключили свет*

Предложение (15) представляется вполне естественным, а предложение (14) вызывает некоторые сомнения. Возникает вопрос, чем предложение (14) хуже, чем (5). Предложения эти отличаются отношением МЕСТА к ВЕЩИ. В (5) МЕСТО так сказать, не зависит от ВЕЩИ. А в (14) и (15) МЕСТО (*дома*), это как бы возможный параметр ВЕЩИ. И для такого рода ВЕЩЕЙ соотношение структуры перспективы с возможностью синхронной или ретроспективной интерпретации нетривиально. В работе Падучева (2006а) рассматриваются такого рода примеры – см. (16):

(16) а. ?*Меня нет дома*; б. *Я не дома*; в. *Меня не было дома*.

К этим примерам можно добавить главное предложение примера (14) *Я не был дома*, для которого синхронная интерепретация не очень естественна.

Список литературы

1. Апресян Ю.Д. 1980. Типы информации для поверхностно-семантического компонента модели «Смысл ⇔ Текст»//Wiener Slawistischer Almanach. Wien.: Sonderband № 1, 1980. (Перепечатано в Ю.Д. Апресян. *Избранные труды* Том II// М.: 1995. С. 8-101).
2. Апресян Ю. Д.. Лексикографические портреты (на примере глагола *быть*) // Сб. Научно-техническая информация. М.: 1992. Серия 2, № 3. (Перепечатано в Ю.Д. Апресян. Избранные труды. Том II// М.: 1995. С. 503-534).
3. Апресян Ю. Д. О Московской семантической школе// Вопросы языкознания. М.: 2005. № 1, С.3-31.

010-Borschew_V:_Layout 1 18.05.2008 14:21 Page 66

4. Борщев В.Б. и Барбара Х. Парти. О семантике бытийных предложений// Семиотика и информатика. М.: ВИНИТИ, 2002. Вып. 37, С. 59-78.
5. Борщев В.Б. и Барбара Х. Парти. Бытийные, локативные и другие близкие к ним предложения// В печати.
6. Падучева Е.В. Наблюдатель: типология и возможные трактовки// Труды международной конференции ДИАЛОГ. М.: 2006a. С.403-413.
7. Падучева Е.В. Родительный отрицания и проблема единства дейктического центра высказывания// Известия РАН. Серия литературы и языка. М.: 2006b. № 4, С.3-9.
8. Падучева Е.В. Дискурсивные слова и категории: режимы интерпретации// В. сб. «Исследования по теории грамматики». М.: Гнозис, 2008. Вып. 4, С.28-57.
9. Падучева Е.В. Бытие и восприятие в семантике генитивной конструкции отрицания// Рукопись 2007 http://www.lexicograph.ru/news/22/
10. Paducheva Elena V. In print. Locative and existential meaning of the Russian *byt'* // *Russian Linguistics*. http://www.lexicograph.ru/news/22/

011-Braslavsky_P:_Layout 1 18.05.2008 14:44 Page 67

СРАВНЕНИЕ ПЯТИ МЕТОДОВ ИЗВЛЕЧЕНИЯ ТЕРМИНОВ ПРОИЗВОЛЬНОЙ ДЛИНЫ

COMPARISON OF FIVE METHODS FOR VARIABLE LENGTH TERM EXTRACTION

Браславский П.И. *(pb@imach.uran.ru),* ***Соколов Е.А.*** *(esokolov@list.ru)*
Институт машиноведения УрО РАН, Екатеринбург

В статье рассматриваются пять методов автоматического выделения и «сборки» терминоподобных конструкций произвольной длины. Проведены эксперименты на корпусе статей «Информационного вестника ВОГиС». Предложена комбинированная методика оценки (экспертная и формальная), приведены результаты сравнительной оценки методов.

1. Введение

Задача выделения ключевых слов и терминов из текста возникает в библиотечном деле, лексикографии и терминоведении, а также в информационном поиске. Объемы и динамика информации, которая подлежит обработке в этих областях в настоящее время, делают особенно актуальной задачу *автоматического выделения* терминов и ключевых слов. Выделенные таким образом слова и словосочетания могут использоваться для создания и развития терминологических ресурсов, а также для эффективной обработки документов: индексирования, реферирования, классификации.

В наших предыдущих работах [2, 1] мы исследовали методы автоматического выделения *двухсловных* терминов-кандидатов из текста. В работе [2] мы сравнивали методы выделения устойчивых словосочетаний, которые используют: 1) статистику встречаемости пар и отдельных слов в тексте и 2) морфологические шаблоны-фильтры. Мы сравнили четыре метода: 1) прямой подсчет количества пар (**freq**); 2) **t-тест**; 3) **χ^2-тест**; 4) отношение функций правдоподобия (**LR**). Как показала оценка, методы **freq** и **t-тест** сравнимы по эффективности и могут быть использованы для составления списка терминов-кандидатов в задачах полуавтоматического формирования терминологических ресурсов. Основной тип ошибок обоих методов – выделение устойчивых общеупотребительных словосочетаний, удовлетворяющих шаблонам. В работе [1] для повышения точности выделения терминов мы предложили использовать Веб в качестве контрастного корпуса, доступ к которому осуществляется с помощью поисковых машин интернета. Для отделения терминоподобных словосочетаний от общеупотребительных выражений мы использовали два параметра: 1) частотность словосочетания и 2) совместная встречаемость словосочетаний. Использование статистики по Вебу позволило улучшить качество выделения двухсловных терминов из корпуса статей «Информационного вестника ВОГиС».

В данной работе мы исследуем различные методы, которые могут быть использованы для выделения терминоподобных словосочетаний произвольной длины и структуры. Сложность этой проблемы в том, что для ее решения статистические подходы не так эффективны: при увеличении длины термина падает частота его встречаемости, в специализированном корпусе ограниченного объема термин может встречаться один-два раза. Наша первоначальная идея состояла в том, чтобы оценить возможность (и сложность) выделения длинных терминов, которые могут встречаться редко (даже один раз) в исследуемом тексте/корпусе. Поэтому методы характеризуются высокой полнотой и – как следствие – низкой точностью выделения терминов. Коль скоро ставится задача выделения терминов *произвольной* длины, делается минимум предположений о структуре термина (при реализации методов мы ввели ограничение: термины могут состоять только из существительных, полных прилагательных, причастий и порядковых числительных). Наш подход можно назвать подходом «чистой доски» (*knowledge-poor approach*): на этапе выделения терминов-кандидатов мы используем минимум информации о структуре и составе терминов, не используем словари, тезаурусы и другие *семантические* ресурсы, не делаем привязки к определенной предметной области. В процессе проведения эксперимента мы скорректировали наш план и провели автоматическую оценку методов с учетом частоты встречаемости кандидатов в термины (к сожалению, у нас не было возможности повторно провести экспертную оценку с учетом частоты встречаемости).

011-Braslavsky_P:_Layout 1 18.05.2008 14:44 Page 68

2. Исследуемые методы

Для сравнительного анализа мы выбрали и реализовали пять методов выделения терминов произвольной структуры.

2.1 MaxLen

Статья [8] описывает одну из первых систем для автоматизированного извлечения терминологии. Система LEXTER выделяет термины из корпуса технических текстов на французском для последующей обработки экспертом. Первый этап работы системы – выделение максимальных цепочек, содержащих термины. Эти цепочки определяются негативно: составляется список слов и знаков, которые *не могут* входить в термин. В нашей реализации в качестве таких разделителей мы рассматриваем знаки препинания, стоп-слова, глаголы, деепричастия; строки между этими разделителя рассматриваются как кандидаты в термины. Это наиболее простой из рассматриваемых методов.

2.2 C-value

Метод выделения многословных терминов, предложенный Frantzi et al. [9], поощряет словосочетания, не входящие в состав других, более длинных. Встречаемость длинных терминов в тексте ниже, чем коротких, и метод ***C-value*** был предложен для компенсации этого эффекта. Значение терминологичности рассчитывается так:

$$C-Value(a)=\begin{cases}\log_2|a|*freq(a), если\quad не\quad вложен\\ \log_2|a|*freq(a)-\dfrac{1}{P(T_a)}*\sum_{b\in T_a}freq(b)\end{cases},$$

где

a – кандидат в термины,

$|a|$ - длина словосочетания, измеряемая в количестве слов,

$freq(a)$ – частотность a,

T_a – множество словосочетаний, которые содержат a,

$P(T_a)$ – количество словосочетаний, содержащих a.

Легко видеть, что чем больше частота термина-кандидата и его длина, тем больше его вес. Но если этот кандидат входит в большое количество других словосочетаний, то его вес уменьшается.

2.3 k-factor

Метод, который мы обозначили ***k-factor***, предложен в работе [7] и реализован в системе BootCaT. BootCaT служит для автоматического формирования тематического корпуса из Веба. Построение корпуса начинается с набора исходных терминов (*seed terms*). С помощью автоматических запросов к поисковой машине извлекаются документы, содержащие исходные термины; в свою очередь из этих документов извлекаются новые *однословные* термины (на основе сравнения частот в сформированном корпусе со «стандартным корпусом»), которые вновь можно использовать в качестве запросов, и т.д. Финальный корпус и список однословных терминов используется для итеративного извлечения многословных терминов. Метод можно рассматривать как упрощенный вариант метода ***C-value***: если более короткий термин-кандидат встречается лишь немногим чаще, чем более длинный термин-кандидат, в который он полностью входит, то «основным» считается более длинный вариант. Отбором управляет пороговое значение отношения частот терминов k (в нашей реализации, как и в [7], k=0,7).

2.4 Window

Метод, описанный в [3], мы условно обозначили ***Window*** (в оригинальной статье он обозначен TERMS—). Идея метода близка двум предыдущим (***C-value, k-factor***) – наращивать словосочетания, если более короткие часто встречаются в составе более длинных. Однако в отличие от других методов, учитывается не только частота контактных случаев (слова непосредственно следуют друг за другом), но и совместная встречаемость в *окне*. На каждой итерации для каждого элемента списка запоминается его непосредственные соседи и соседи в текстовом окне. Создаются соответствующие таблицы, вычисляется частотность встречаемости пар. Далее, предполагается, что если пара элементов (на первом этапе – отдельных слов) встречается как непосредственные соседи более чем в половине случаев их появления в одном и том же текстовом окне, то эта пара представляет собой термин или фрагмент термина. Происходит склейка пары в единый элемент, таблицы пересчитываются

011-Braslavsky_P:_Layout 1 18.05.2008 14:44 Page 69

так, как если бы этот элемент был известен с самого начала, до начала обработки текста, что дает возможность и дальше наращивать термин. Авторы приводят примеры длинных терминов, полученных этим методом: *закон об обязательном страховании гражданской ответственности владельцев транспортных средств*[1], *исполнительный орган местного самоуправления*. В нашей реализации размер окна – 9 слов. Если не накладывать ограничений на частоту встречаемости склеиваемых элементов, то метод объединит уникальные (с частотой 1) цепочки допустимых слов (т.е. повторит результат ***MaxLen***).

2.5 Синтаксический анализ (AOT)

Известно, что большинство терминов – это именные группы (хотя в [6], например, показано, что номинативность не является исключительной характеристикой терминов во многих предметных областях). В рамках этого метода в качестве терминов-кандидатов мы рассматриваем именные группы, выделенные с помощью синтаксического анализатора. Метод получил название по используемому анализатору – ***AOT*** [5]. В нашей реализации мы брали синтаксические группы ПРИЛ_СУЩ и ГЕНЕТ_ИГ. После первичного анализа мы проредили список, исключив группы с однородными рядами (с запятыми, союзами *и/или*), а также группы, не содержащие ни одного русского слова (обычно библиографические ссылки). Полученные строки мы преобразовывали так, чтобы главное слово именной группы было словарной форме; другой обработки не проводилось (мы никак не обрабатывали стоп-слова, поэтому, например, самое частотное из выделенных методом словосочетаний – *то же время*). Метод на основе AOT – единственный из рассматриваемых, который допускает наличие предлогов в составе кандидатов в термины.

3. Данные и инструменты

Мы проводили эксперименты на корпусе статей «Информационный вестник Вавиловского общества генетиков и селекционеров (ВОГиС)» [4], который использовался в наших предыдущих экспериментах [1]. Корпус содержит 100 статей разных авторов по генетике, селекции, а также смежным наукам, опубликованных в «Информационном вестнике ВОГиС» с 1997 по 2006 год. Были взяты все статьи журнала за этот период, за исключением редакционных статей, посвященных юбилеям ученых и памятным датам. Характеристики корпуса: всего слов – 256 255, без стоп-слов – 179 635. Система *AOT* выделила в корпусе 35 737 предложений (судя по результатам, правила для конца предложения довольно простые: так, комбинация «точка+пробел» всегда интерпретируется как конец предложения), *mystem* выделил 27 880 предложений.

Для формальной оценки результатов мы используем русскую часть словаря терминов по молекулярной и клеточной биологии (http://www.mblogic.net/glossary/), предоставленную нам Анастасией Барышниковой. Словарь содержит примерно 6 300 входов (строк). Каждая строка может содержать несколько близких терминов, например: *гипотеза гибридной ДНК; модель гибридной ДНК; гетеродуплексная модель ДНК; полярон-гибридная модель ДНК*. Мы рассматриваем все термины словаря как равноправные (всего – 7 199), распределение длин терминов словаря выглядит следующим образом: 1 слово – 2 941; 2 слова – 3 110; 3 слова – 798; 4 слова – 214; 5 и больше слов – 136.

Интересно отметить, что словарь содержит достаточно много «терминов-метафор» (обычно употребляются в кавычках) – как однословных (*«аркан», «булава», «восьмерка», «газон»* и др.), так и многословных (*«шитье назад», «горячая точка», «узлы-на-веревке», «счастливые уроды»* и др.). Кроме того, словарь содержит много терминов специфической структуры, например, с цифрами (*1-метил-4-амино-6-оксипиримидин, 4-тиоуридин* и др.), греческими и латинскими буквами (*α-гетерохроматин*, *β-талассемия*, *D-петли, F-эписома, HKG-бэндинг* и др.), а также сложные термины (например, *хронический остеомиелит длинных костей после огнестрельных повреждений*).

Тексты анализировались в формате *plain text*. Корпус обрабатывался как монолитный документ, без учета разбиения на отдельные статьи. Морфологическая обработка (кроме метода *AOT*) осуществлялась с помощью программы *mystem* (http://company.yandex.ru/technology/products/mystem/mystem.xml).

4. Методика оценки

Как и в предыдущих работах, мы комбинируем ручную (экспертную) оценку и формальную оценку по «эталонному списку» (словарю). В данном эксперименте мы несколько модифицировали методику, описанную в [2, 1].

В соответствии с нашим первоначальным планом, для экспертной оценки мы брали по 100 кандидатов из

[1] В нашей реализации этого метода предлоги не могут входить в состав строки-кандидата.

011-Braslavsky_P:_Layout 1 18.05.2008 14:44 Page 70

полных результатов работы каждого метода («длинного списка»). Для ***C-value*** мы брали верхушку (top100) отсортированного списка, для остальных методов – 100 случайных строк из списка с учетом длин: 33 – трехсловных, 33 – четырехсловных, 34 – длины пять и более слов. Объединенный «короткий» список содержит 492 строки.

Экспертная оценка организована следующим образом. Сначала эксперту предъявляется краткое описание предметной области, а также несколько положительных и отрицательных примеров терминов для данной области. После этого эксперт, используя простой интерфейс, последовательно для каждого элемента списка отвечает на вопрос: «Является ли данное словосочетание термином предметной области?» Варианты ответа эксперта: «да», «нет», «затрудняюсь ответить», а также «частично» (предъявленное словосочетание содержит термин или является частью более длинного термина). Список предъявляется эксперту «порциями» по 10 словосочетаний, порядок предъявления словосочетаний – случайный. Каждый термин-кандидат оценивается независимо двумя экспертами в данной предметной области. В случае *сильной оценки* термином считается словосочетание, которое оба эксперта признали термином; в случае *слабой оценки* только один из экспертов оценил словосочетание как термин.

Формальную оценку мы провели для «короткого», «длинного» и «среднего» списков. В последний попали только строки из «длинного списка» с частотой встречаемости больше единицы. Мы проводим два типа формальной оценки на основе 1) четкого и 2) нечеткого сравнения. В первом случае мы подсчитываем три параметра: 1) *точные совпадения* выделенных терминов с терминами словаря, 2) *включение* терминов словаря в выделенные словосочетания и 3) *вхождение* выделенного словосочетания в более сложные (четыре и более слова) термины словаря. При *нечеткой* оценке мы рассматриваем словосочетания как множества слов, приведенные к нормальной форме, а близость двух строк определяем как отношение количества совпавших слов к общему количеству уникальных слов в двух словосочетаниях: $sim\ (S_1, S_2) = |S_1 \cap S_2| / |S_1 \cup S_2|$. При оценке мы считаем количество терминов-кандидатов, для которых в словаре есть хотя бы один термин с близостью ≥0.5. Примеры близких терминов по этой метрике приведены в табл. 1.

5. Результаты

Примеры строк, которые были переданы экспертам для оценки, представлены в табл. 2. Результаты оценки списка из 492 кандидатов в термины («короткий список») приведены в табл. 3. Оценки экспертов совпали в 54% случаев. В рамках нашего предыдущего эксперимента мнения тех же экспертов при оценке двухсловных кандидатов из того же корпуса совпади почти в 80% случаев [1]. Очевидно, что причина двоякая: длинные цепочки слов менее однозначны и устойчивы, к тому же в данном эксперименте у экспертов было больше вариантов оценки (четыре против трех в предыдущем эксперименте). На 27 оценках (5%) мнения экспертов противоположны: «термин» vs. «не термин».

В табл. 4 приведены результаты формальной оценки полных результатов («длинный список») обработки корпуса ВОГиС различными методами. В табл. 5 сведены результаты формальной оценки «среднего списка» (состоит из строк, которые встречаются в корпусе минимум два раза), в табл. 6 приведены результаты нечеткой оценки кандидатов с учетом их длины в словах.

Графики на рис. 1, 2 построены на основе обработки списков строк, полученных в результате работы разных методов и упорядоченных по частоте встречаемости в корпусе. Рис. 1 соответствует top500 всех пяти методов, рис. 2 – «среднему списку», полученному с помощью метода *C-value* (2 466 строк).

Строка-кандидат	Термин из словаря	sim
подавление экспрессии гена	экспрессия гена	0.67
стволовая нервная клетка	стволовая клетка	0.67
центральная нервная система	вегетативная нервная система	0.5
центральная нервная система	центральная нервная система	1.0
подавление экспрессии гена мишени	экспрессия гена	0.5
институт химической биологии	институт химической биологии и фундаментальной медицины	0.5
действие естественного отбора	естественный отбор	0.67
полимеразная цепная реакция	полимеразная цепная реакция	1.0
полимеразная цепная реакция	обратная полимеразная цепная реакция	0.75
российская академия наук	российская академия медицинских наук	0.75
генетическая дифференциация популяции	генетическая структура популяции	0.5
фенотип множественной лекарственной устойчивости	множественная устойчивость к лекарственным препаратам	0.5

Таблица 1. Примеры вычисления нечеткой близости строк

011-Braslavsky_P:_Layout 1 18.05.2008 14:44 Page 71

MaxLen
уникальный цветовой баркод
боковые передающие цепочки
показатели ассортативности браков
фуражная ценность зеленой массы
здоровье населения алтайского края
уникальный генофонд пушных зверей
средним величинам антропометрических признаков метисы
сущность процессов редукционной эволюции субгеномов органелл
разнообразного типа специализированные дифференцированные клетки
геологами докембрийская летопись развития органического мира Земли
C-value
подавление экспрессии генов
теория естественного отбора
отрицательная обратная связь
центральная нервная система
расщепление фосфодиэфирной связи
экспрессия гена мишени
вторичная структура рнк
регуляция экспрессии гена
боковая петля рнк
множественная лекарственная устойчивость
подавление экспрессии гена мишени
k-factor
нехватка ионов марганца
Ненецкий автономный округ
резкое усиление действия
ядра клеток вентральной нейроэктодермы
часть указанных сложных вопросов
главный позитивный эффект миграции
частота наследственных болезней человека
период высокие индексы брачной ассортативности
кинетические характеристики расщепления синтетических фрагментов
нарушения общей устойчивости обмена веществ
Window
мощный способ идентификации
результат филогенетического анализа
уровень генетического разнообразия
наибольшее разнообразие микросателлитных гаплотипов
общие проблемы физико-биологии
упорный поиск общих законов наследования
отсутствие содержательной интерпретации анализируемых признаков
заманчивая легенда получения зерновой культуры
экспериментальная проверка эффективности селекционного индекса
генетическая характеристика удэгейцев Приморского края
AOT
значимые коридоры миграции
различные гибридные комбинации
обоснованное хирургическое вмешательство
более древняя история происхождения
судьба больших групп животных
двигательная активность неонатальных крысят
благо совместных творческих междисциплинарных исследований
генетическая история алеутов Командорских островов
резкое снижение численности коренного населения
2000 полных митохондриальных геномов индивидуумов различного этнорасового происхождения

Таблица 2. Примеры выделенных строк

011-Braslavsky_P:_Layout 1 18.05.2008 14:44 Page 72

Оценка		*MaxLen*	*C-value*	*k-factor*	*Window*	*AOT*
Экспертная, «термин»	слабая	30	62	30	25	20
	строгая	8	24	6	2	5
Экспертная, «частично»	слабая	44	38	59	53	47
	строгая	14	7	21	13	12
Формальная	точно	0	5	0	0	0
	включение	66	70	70	71	63
	вхождение	0	6	0	0	0
	нечеткая	6	35	8	8	8

Таблица 3. Результаты оценки «короткого списка»

Оценка	*MaxLen*	*C-value*	*k-factor*	*Window*	*AOT*
размер списка	14 970	34 370	16 986	13 845	18 772
точно	23	34	27	14	33
включение	10 309	23 322	11 663	9 300	10 579
вхождение	11	29	8	6	17
нечеткая	1 613	3 640	1 836	1 382	1 712

Таблица 4. Результаты формальной оценки «длинного списка»

Оценка	*MaxLen*	*C-value*	*k-factor*	*Window*	*AOT*
размер списка	743	2 466	1 949	1 352	1 190
точно	10	20	21	13	18
включение	492	1 643	1309	883	726
вхождение	4	15	5	3	6
нечеткая	150	501	420	267	196

Таблица 5. Результаты формальной оценки «среднего списка»

Длина	*MaxLen*		*C-value*		*k-factor*		*Window*		*AOT*	
	всего	близко	всего	близко	всего	близко	всего	близко	всего	близко
3 слова	597	118	1 963	409	1 609	342	1 075	208	987	160
4 слова	114	27	375	82	273	69	230	54	171	34
5 и больше	32	5	128	10	67	9	47	5	32	2
Всего	743	150	2 466	501	1 949	420	1 352	267	1 190	196

Таблица 6. Результаты нечеткой формальной оценки «среднего списка» с учетом длины кандидатов в термины

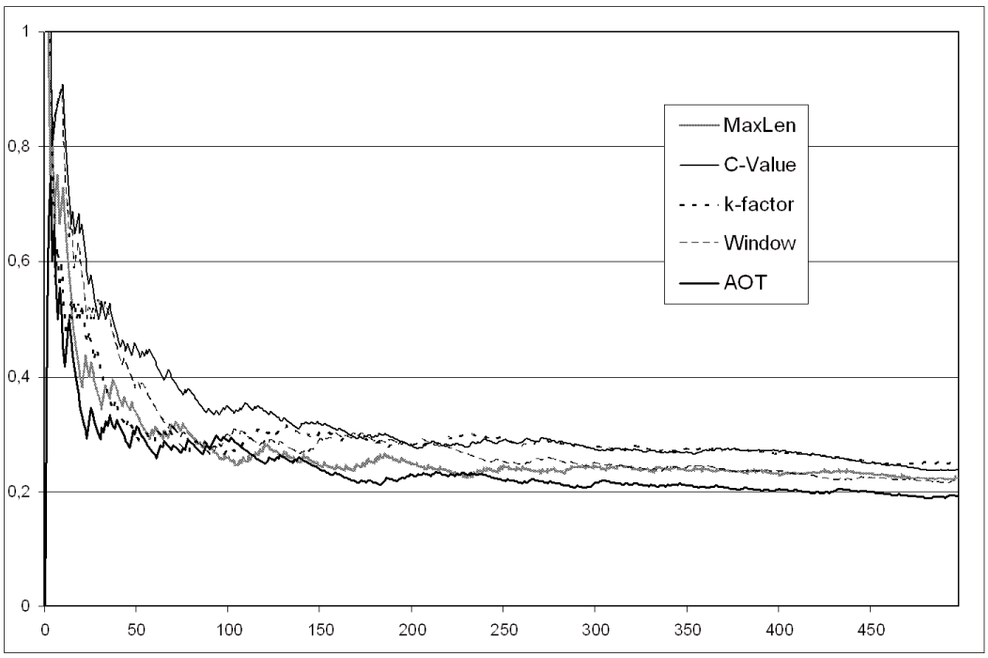


Рис. 1. Доля строк, близких словарным, в зависимости от длины списка. Top500 «среднего списка», упорядоченного по убыванию частот

011-Braslavsky_P:_Layout 1 18.05.2008 14:44 Page 73

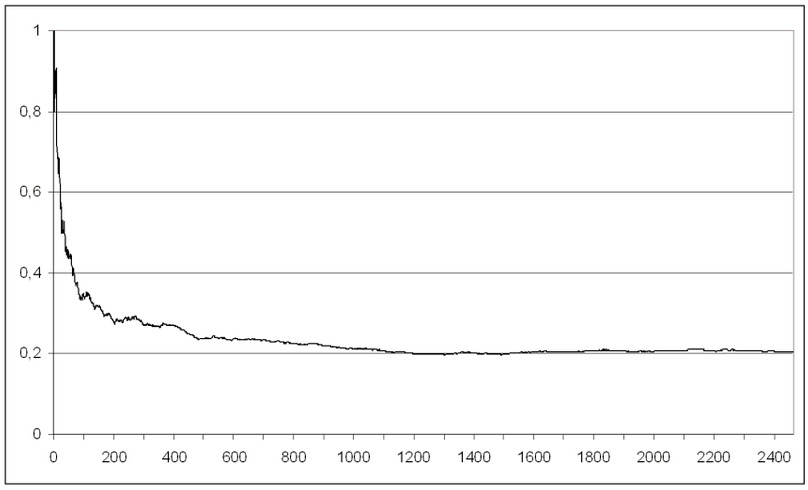


Рис. 2. Доля строк, близких словарным, в зависимости от длины списка. «Средний список» метода C-value, упорядоченный по убыванию частот (2 466 строк)

6. Выводы

На основе анализа результатов можно сделать вывод, что сравниваемые методы дают в целом похожие результаты. Несколько лучше ведут себя методы, учитывающие вложенность терминов (*C-value, k-factor*). Выделение именных групп на основе синтаксического анализа без дополнительных ограничений дает худший результат (*AOT*). Сопоставление результатов экспертной и формальной оценок (см. табл. 3) позволяет сделать вывод, что формальные методы годятся для сравнения больших списков кандидатов в термины.

Учет частоты встречаемости строк существенно не повышает качество выделения терминов, если нас интересует хоть сколько-нибудь значительная полнота: на основании формальной оценки (см. рис. 1) можно предположить, что точность деградирует очень быстро (во всяком случае, для небольшого корпуса текстов). При этом в области редких строк «всегда есть что-то интересное»: примерно для каждой пятой строки на большом диапазоне списка есть близкий словарный термин почти постоянна (см. рис. 2).

Распределение длин терминов из словаря по молекулярной и клеточной биологии, а также сильное ухудшение качества выделения длинных терминов (5 и более слов) подсказывает, что повысить качество и обеспечить хорошую полноту можно с помощью шаблонов для трех- и четырехсловных терминов. Как предлагается в других работах (например, [3]), использование семантических словарей сочетаемости, продуктивных и непродуктивных слов может существенно повысить качество выделения и сборки терминов.

Благодарности

Мы благодарим Анастасию Барышникову и Татьяну Струкову, которые приняли участие в оценке. Кроме того, мы благодарим Анастасию за предоставленный словарь, а также анонимного рецензента за ценные замечания по содержанию работы.

Список литературы

1. Браславский П., Соколов Е. Автоматическое извлечение терминологии с использованием поисковых машин интернета // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. Диалог‘2007. М.: Изд-во РГГУ, 2007. С. 89-94.

2. Браславский П., Соколов Е. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. Диалог’2006. М.: Изд-во РГГУ, 2006. С. 88–94.

3. Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических сочетаний по текстам предметной области // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды пятой Всероссийской научной конференции (С.-Петербург, 29-31 октября 2003 г.), 2003. С. 201–210.

4. Информационный вестник ВОГиС, http://www.bionet.nsc.ru/vogis/

011-Braslavsky_P:_Layout 1 18.05.2008 14:44 Page 74

5. Синтаксический анализ. Проект AOT, http://www.aot.ru/docs/synan.html

6. Шелов С.Д. Терминоведение: семь вопросов и семь ответов по семантике термина // НТИ. Сер. 2. Информационные процессы и системы, 2001. №2. С. 1-11.

7. Baroni M., Bernardini S. BootCaT: Bootstrapping Corpora and Terms from the Web // Proceedings of LREC 2004. Lisbon: ELDA, 2004. P. 1313–1316.

8. Bourigault D. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases // Proc. of COLING-92, Nantes, France, August 23-28, 1992. P. 977–981.

9. Frantzi K., Ananiadou S., Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method // Int J Digit Libr (2000) 3: 115–130.

012-Buzikashvili_H:_Layout 1 18.05.2008 14:40 Page 75

МНОГОЗАДАЧНЫЙ ПОИСК: ФАКТ, АРТЕФАКТ, ПРЕНЕБРЕЖИМОЕ ИСКЛЮЧЕНИЕ?

MULTITASKING SEARCH: FACT, ARTIFACT, OR NEGLIGIBLE EXCEPTION?

Бузикашвили Н.Е. *(buzik@cs.isa.ru)*
Институт системного анализа РАН

Рассмотрено, как часто и в какой форме пользователь поисковой машины Интернета выполняет несколько поисковых задач параллельно. Показано, что параллельное исполнение очень редко и ограничивается двумя задачами, выполняемыми как вложенные. Характеристики временных сессий, содержащих одну, несколько последовательно и несколько параллельно исполняемых задач, различны, а наиболее простым из трех классов оказались последовательно исполняемые задачи.

1. Введение

Практические ситуации, в которых до сих пор изучалась многозадачность – это, прежде всего, ситуации *вынужденной* многозадачности, когда, оператор, в силу внешних обстоятельств, принужден контролировать сразу несколько процессов [6,9]. Типичный пример – работа диспетчера. Принципиальное отличие *невынужденной* многозадачности – принятие самим оператором решения о параллельном исполнении задач. Это решение может быть связано с ситуативной оценкой ожидаемых затрат и рисков, а может носить "ритуальный" характер или иметь форму автоматизма.

Большинство лабораторных исследований показывает, что многозадачность более затратна и менее эффективна, чем последовательное выполнение тех же задач [6,7,9], и невынужденная многозадачность априори должна быть редкой. Поиск в Интернете относится как раз к этой категории. Поэтому, выводы [11–14] о массовом характере многозадачности, согласно которым многозадачными являются 11-12% поисковых сессий пользователей Интернета, достаточно неожиданны. С другой стороны, легко найти (пусть и менее частые) ситуации, когда многозадачность хотя и более затратна, но *приемлема* для исполнителя, либо когда затраты на параллельное исполнение задач *не выше*, чем на последовательное. Поэтому результаты [11–14] могут быть верны. В таком случае, нужно отнести поиск к числу немногих сфер, где многозадачность приемлема.

В [11–14] разметка запросов пользователя на задачные сессии выполнялась вручную, а критерий отнесения разных запросов к одной задачной сессии не указан. Поэтому вероятная ценность [11–14] ограничивается самим фактом обнаружения многозадачности, использованный же метод (ручная разметка) едва ли представляет интерес. Если многозадачность редка, необходимо использовать автоматическую или, как минимум, полуавтоматическую процедуру. Ее суть состоит в том, чтобы на первом, автоматически выполняемом шаге исключить из дальнейшей обработки сессии, почти наверняка не являющиеся многозадачными, после чего асессор выбрал бы среди сессий, автоматически опознанных как многозадачные, те, что действительно являются ими. Сессии, автоматически не опознанные как многозадачные, асессору не предъявляются. Такая процедура эффективна, только если на автоматическом шаге *ошибка отбраковки* (опознание многозадачной сессии как немногозадачной) мала, а доля отбракованных сессий – велика. Последнее необходимо для многократного сокращения затрат асессора. Величина *ошибки неотбраковки* (ложного опознания сессии как многозадачной) не важна, т.к. эти ошибки, в отличие от ошибок отбраковки, устраняются на этапе ручной проверки.

Построенная нами автоматическая процедура отбраковывала как не-многозадачные порядка 99% всех временных сессий. На этапе выборочной ручной проверки только чуть более половины из них были признаны действительно многозадачными.

Были сформулированы гипотезы и вопросы о том, как пользователь выполняет многозадачные сессии (сколько задач может исполняться параллельно, как чередуется их исполнение, как возобновляются прерванные задачи), которые были проверены автоматически, на полных наборах данных. Ручной контроль подтвердил выводы, полученные при автоматической обработке.

Кроме того, рассмотрены числовые характеристики трех классов сессий: (*а*) однозадачные сессии, (*б*) линейные, в которых задачи исполняются последовательно одна за другой, и (*в*) собственно многозадачные, в

012-Buzikashvili_H:_Layout 1 18.05.2008 14:40 Page 76

которых одни задачи временно прерываются для исполнения других. Оказалось, что характеристики этих классов различны. По их значениям можно судить о сложности исполнения каждого из классов сессий. То, что многозадачные сессии наиболее трудоемки, не удивительно. Но также оказалось, что сложность задачи в однозадачной сессии выше, чем у задачи в линейной сессии.

2. Многозадачность: исполнительные цепочки, разделяемые ресурсы

Термин “многозадачный поиск” применительно к поисковому поведению пользователя Интернета был впервые использован А.Спинк [11–14]. Что понимают под многозадачностью в смежных с исследованием поискового поведения областях? Почти полвека назад “многозадачный режимом” была названа такая работа операционной системы, при которой несколько задач выполняются “одновременно”. Операционные системы работали на однопроцессорных машинах, и многозадачный режим был синонимом режима с разделением времени, при котором несколько задач, чередуясь, использовали общий ресурс. Совершенно в этом же смысле многозадачность используется в когнитивной и инженерной психологии. Последовательное выполнение задач не называется многозадачностью и содержательно эквивалентно однозадачности.

Говоря о многозадачности, имеют в виду одну из двух форм: (*а*) исполнение разных задач в один и тот же момент времени или (*б*) чередование задач по схеме прерывание задачи – выполнение другой – ее прерывание и возобновление ранее прерванной.

Компьютерная (=физиологическая) метафора делают вопрос о многозадачности совершенно прозрачным и позволяют обойти неплодотворную психологическую дискуссию “многозадачность – сложнее или проще?”. Есть набор *ресурсов* (процессоры, разделы мозга, периферийные устройства и органы), используемых в задачах. Задаче соответствует *исполняющая цепочка* – последовательность используемых ею ресурсов. Если цепочки не пересекаются, то задачи могут исполняться одновременно, что не ведет к увеличению затрат. Однако если же две цепочки используют один и тот же ресурс, то возникает проблема разделения этого ресурса. Именно ситуации пересечения цепочек типичны для лабораторных исследований многозадачности. Заметим, что и в этом случае поведение оператора скорее описывается не критерием минимизации затрат, а критерием их приемлемости, т.е. не превышают ли они некоторый порог.

Одновременное исполнение непересекающихся задач, когда оно возможно, часто проще и приятнее чередования. Напротив, пересечение (обычно совпадение) исполняющих цепочек может сделать совокупность относительно простых задач невыполнимой, как например, в сеансе одновременной игры вслепую.

В терминах исполняющих цепочек, многозадачный поиск – это тотальное разделение ресурсов: все поисковые задачи используют одни и те же ресурсы, а несовпадение исполняющих цепочек возможно только в конечном звене – окне поиска. Поскольку современная поисковая среда обеспечивает возможность выполнять несколько задач параллельно в разных окнах, препятствием поисковой многозадачности могут быть только ресурсы самого пользователя, но не поисковый интерфейс.

Воспользуемся еще одной, приблизительной метафорой. Когда человек несколько раз подбросит шарик ○, а затем несколько раз второй ●, мы не говорим о жонглировании, а затраты здесь те же, как при подбрасывании одного шарика. Однако такое “псевдожонглирование” есть точный аналог последовательного исполнения нескольких задач. Каков в действительности порядок выполнения транзакций разных задач при их параллельном исполнении? Случайный, регулярный и если регулярный, то что это за регулярность. Сколькими задачами может одновременно “жонглировать” пользователь? На Рис. 1 изображены все варианты “жонглирования” двумя поисковыми задачами, содержащими 2 и 3 транзакции. Определяющее сложность исполнения число возобновлений незавершенной задачи равно 0 для перестановок, соответствующих последовательному исполнению. Следующее по сложности – вложенное исполнение, в котором одна задача, прервав другую, исполняется до конца, и самое сложное – смена задачи после каждой транзакции (“истинное жонглирование”).

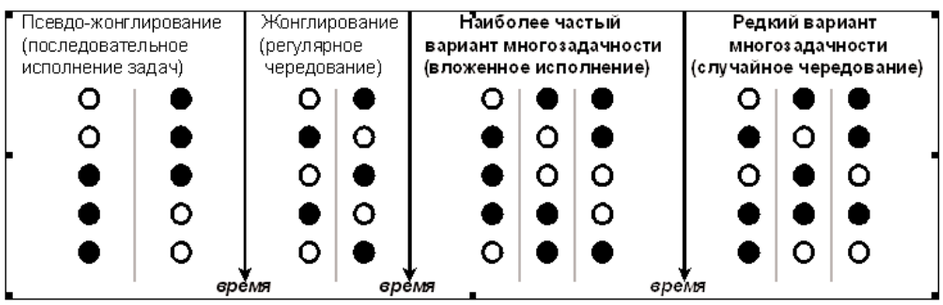


Рис. 1. “Жонглирование” двумя задачами

012-Buzikashvili_H:_Layout 1 18.05.2008 14:40 Page 77

3. Определения

В статье используются следующие понятия:

(1) *транзакция* – акт взаимодействия пользователя с поисковой машиной. Транзакция может соответствовать либо *подаче запроса*, либо *листанию* страницы (экрана) результатов запроса. Для описания транзакции будем использовать пару (*запрос, страница_результатов*). Например, если пользователь (1) подал запрос и получил нулевую страницу результатов (1-я транзакция), (2) перелистнул результаты на следующую страницу (2-я транзакция), (3) вернулся к нулевой странице (3-я транзакция), то последовательность описаний транзакций: (*запрос*, 0), (*запрос*, 1), (*запрос*, 0).

(2) *временная сессия* – последовательность транзакций одного пользователя с поисковой машиной, отделенная от предыдущих и последующих его транзакций временем, не меньшим используемым в исследовании *межсессионным временным порогом*. В разных исследованиях этот порог варьируется от 5 мин [10] до 2 часов [8]. В работах [11–14] использовалось 15-минутное значение.

(3) *задача* – поиск одного и того же объекта или очень сходных объектов;

(4) *задачная сессия* – все транзакции временной сессии, относящиеся к одной задаче. Ниже, для краткости и во избежание путаницы с временной сессией, будем называть задачей именно задачную сессию (что путаницы не повлечет).

(5) рассматриваемые классы временных сессий:

(5a) *однозадачная сессия* (все запросы временной сессии относятся к одной задаче),

(5b) *линейная сессия* из нескольких задач содержит более одной задачи, которые выполняются одна за другой без прерываний незавершенных задач,

(5c) *многозадачная сессия* содержит несколько задач, среди которых хотя бы одна прерывается другой.

(6) Текущая ширина временной сессии – число незавершенных (первая транзакция уже выполнена, а последняя – еще нет) на данный момент (на момент очередной транзакции) задач. *Шириной сессии* назовем ее максимальную текущую ширину.

4. Вопросы и гипотезы

В исследовании мы рассмотрели следующие вопросы и гипотезы (см. также [2,3]):

(*a*) Сколь часто различные задачи выполняются параллельно, какова доля временных сессий, содержащих участки параллельного выполнения двух и более задач?

(*b*) Каково соотношение двух типов возобновления прерванной задачи, т.е. как часто задача возобновляется: 1) вводом нового запроса (точнее, транзакции с 0-й страницей результатов) или 2) листанием результатов запроса, введенного до прерывания?

Гипотеза о возобновлении состоит в том, что поскольку продолжение работы с результатами запроса, поданного перед прерыванием, более затратно, чем подача нового запроса, доля возобновлений-листаний меньше доли листаний среди продолжений линейной сессии.

(*c*) *Гипотеза о схемах переключения* между незавершенными задачами. Если имеется несколько незавершенных задач, то за транзакцией одной задачи, скорее всего, последует ее же транзакция. В частности, это значит, что порядок исполнения “остающихся на данный момент” транзакций незавершенных задач отличен от случайной перестановки.

(*d*) Гипотеза о числовых характеристиках поискового поведения: характеристики параллельно исполняемых задач (длина сессии, длина запроса, число транзакций на запрос) отличаются от характеристик задач, исполняемых линейно одна за другой. Значения числовых характеристик можно интерпретировать в терминах сложности задач и сессий. Предположительно, использование нескольких задач в одной временной сессии более вероятно, если задачи простые. При этом параллельное исполнение сложнее линейного исполнения тех же задач.

(*e*) Естественную гипотезу, что запросы параллельных задач сделаны из разных окон, к сожалению, проверить нельзя, т.к. логи не содержат полей *окно* (*экземпляр браузера*).

5. Наборы данных

Мы использовали логи транзакций двух поисковых машин: (1) почти полностью русскоязычный лог *Яндекса* (недельная выборка, 175000 пользователей, 2005); (2) два набора из почти полностью англоязычного лога *Excite* – полный 8-часовой набор, 1999, 537600 пользователей и суточная выборка 2001, 305000 пользователей. Логи *Excite* многократно использовались ранее, в том числе в исследованиях многозадачности

012-Buzikashvili_H:_Layout 1 18.05.2008 14:40 Page 78

[14], и взяты, чтобы, с одной стороны, сравнивать результаты, полученные на одних и тех же данных в разных исследованиях, а с другой – сравнить результаты, полученные на разных данных одним и тем же инструментом.

Автоматическая предобработка каждого из наборов данных состояла в:

— исключении клиентов, предположительно являющихся роботами. С помощью скользящего временного окна [4] исключались клиенты, сделавшие более 7 *разных* запросов в течение одного часа.

— сегментации последовательности транзакций, выполненных одним пользователем (опознанным как не-робот), на временные сессии. Для сопоставимости с результатами [11–14], использовался тот же межсессионный порог (15 мин), что и в [11–14].

6. Метод

Суть метода, который был применен для выделения задач внутри временных сессий одного пользователя, проста:

(0) запросы, присутствующие в транзакциях одной временной сессии, учитываются однократно. Напр., в сессии (*кот*, 0) (*кот*, 1) (*пес*, 0) есть только 2 запроса – *кот* и *пес*.

(1) для каждой пары запросов, относящихся к одной временной сессии, определяется их сходство: если два запроса содержат хотя бы одно общее слово, являющееся основной частью речи (основными считались существительные, глаголы, прилагательные, причастия и флективно совместимые с ними неизвестные слова), то запросы сходны. Например, <*этот кот*> и <*серый кот* > сходны, а <*этот дом*> и <*этот кот*> – нет.

Однако в запросах, с одной стороны, массовы ошибки (пропуски и перестановки букв, набор одной вместо другой, склейки и вставки пробелов), причем значительная часть слов запросов – не словарные. С другой стороны, какие-то слова просто не имеют однозначного написания, например, иноязычные имена и названия. Чтобы покрыть ошибки и неоднозначности, шаг (1) определения сходства запросов был заменен на (1*), основанный на общих подстроках [1]. Похожий и тем же мотивированный метод использован в [5].

(1*) для всех запросов, относящихся к одной временной сессии строятся парные исходному запросу строки: символы исходного запроса переводятся в строчные, пробелы и пунктуации отбрасываются. Так, запросу <*Жозеф-Луи +Гей-Люссак*> будет сопоставлена строка *жозефлуигейлюссак*. Для каждой пары таких строк определяется бинарное орфографичекое сходство. Именно, если две строки содержат одинаковые подстроки не короче *l*=3 символов, то мы пытаемся посимвольно нарастить их в каждой строке слева и справа с тем, чтобы выделить сходные подстроки. При наращивании допускается 4 типа замен: перестановка соседних символов, вставка и удаление одного символа, замена символа. Применив исправление к одной из текущих сходных подстрок, мы не имеем права применять его к следующим *thr*=3 символам с этого же края подстроки. Когда наращивание схожих подстрок завершено, смотрим, покрывает ли в исходных запросах хотя бы одна из подстрок хотя бы нефлективную часть хотя бы одного слова, являющегося основной частью речи. Если да, запросы считаются сходными. Напр., запросы <*царствующий дом*> и <*шествующий кот*> несходны, т.к. их общая часть

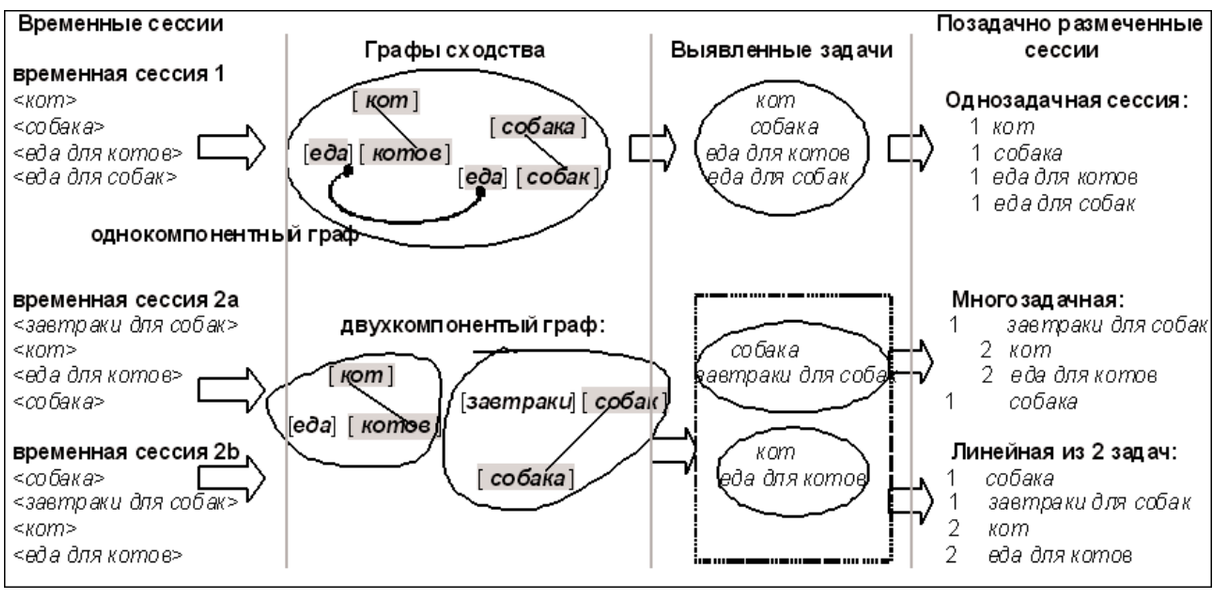


Рис. 2. Примеры выявления задач в трех временных сессиях

012-Buzikashvili_H:_Layout 1 18.05.2008 14:40 Page 79

ствующий приходится на флексию, тогда как запросы <*уфимский богатырь*> и <*сы руфмскйи*> опознаются как сходные. Критерий сходства устойчив к инверсиям, пропускам и вставкам букв, к лишним пробелам, пунктуациям, а также склейкам. Результат шага (1*) – бинарная матрица парного сходства запросов.

(2) После построения матрицы сходства запросов одной временной сессии строится транзитивное замыкание этой матрицы, и запросы, попавшие в одну компоненту связности, считаются относящимися к одной задаче.

(3) Наконец, определяется ширина временной сессии и если сессия широкая, она отписывается в файл сессий данной ширины для последующей ручной проверки.

На Рис. 2 приведен пример работы всей процедуры, а на Рис. 3 – фрагмент результата работы метода выявления задачных сессий (числа слева – маркировка транзакций номерами задач, а положение номеров соответствует уровню вложенности задачи). Как видим, во втором примере (user 119, time session 2) на Рис. 3 сходство ***хлеб****опечения* и *архангельск****хлеба*** не обнаружено, в результате чего ошибочно выявлена задачная сессия 4, в действительности относящаяся к сессии 2.

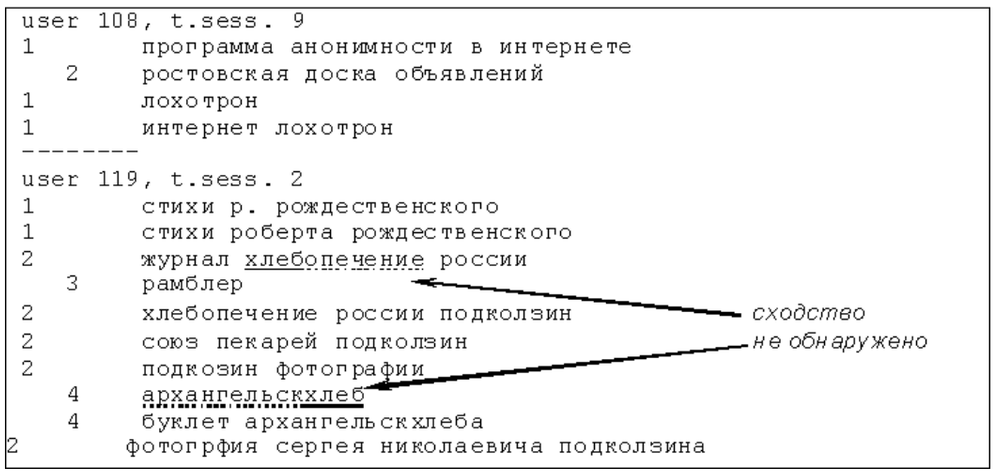


Рис. 3. Фрагмент файла временных сессий ширины 2

Описанная процедура не позволяет выявить сходство орфографически различных, но содержащих синонимы запросов. Однако гораздо чаще классической словарной синонимии в запросах встречается “неклассическая”: (1) “субкультурная синонимия”, т.е. сленг, часто узкогрупповой; (2) “многоязычие”: одно и то же слово дано в запросах то по-русски, то по-английски; (3) ошибки переключения языка; (4) использование наравне с полными формами аббревиатур, а самое интересное – (5) задание не сходных, а *совместимых* признаков, характеризующих разные свойства искомого.

Далее, сходство запросов процедура выявляет систематически реже, а совместимость не выявляет вообще. Это завышает число задач и влечет ложное обнаружение многозадачности. Однако можно воспользоваться систематическим завышением многозадачности как преимуществом.

Именно, после автоматического выявления предположительно многозадачных сессий нужно вручную проверить лишь эти сессии, доля которых составляет порядка 1% (Табл. 1). Т.е. асессор вместо полной разметки 100 сессий должен обработать одну. При этом, при обработке сессии, автоматически уже размеченной на задачи, асессор должен лишь выяснить сходство задач, а не заново разметить все транзакции как в [11–14].

7. Результаты автоматического анализа многозадачного поиска

Некоторые результаты автоматической обработки приведены в Табл. 1. Заметим, что хотя наборы данных относятся к разным языкам, разным поисковым машинам и разным годам, одноименные результаты на всех наборах сходны, что, как надеемся, отражает не влияние метода, а сходство поискового поведения пользователей.

В любой ситуации пользователи выбирают наименее затратную тактику:

(*a*) избегают жонглироть задачами (лишь ~1% временных сессий многозадачные);

(*b*) когда пользователь все же “жонглирует” задачами, он ограничивается 2 “шарами” (в т.ч. когда временная сессия содержит более двух задач), и при этом чаще всего избирает

(*c*) наименее затратный, вложенный вариант многозадачности, при котором число возвратов к прерванной задаче минимально. Именно, прервав одну задачу, он целиком, от начала и до конца выполняет вторую, и лишь

012-Buzikashvili_H:_Layout 1 18.05.2008 14:40 Page 80

завершив ее, возвращается к прерванной. Гипотеза о форме возобновлении не подтвердилась – доля новых запросов как продолжений прерванной задачи та же, что при линейном исполнении, т.е. пользователь часто возвращается к листанию результатов запроса, сделанного перед прерыванием.

(*d*) Число запросов в широких временных сессиях больше, чем в сессиях, в которых задачи исполняются последовательно. Более того, и число листаний результатов одного запроса (т.е. число транзакций на запрос) для широких сессий больше. В то же время запросы широких временных сессий короче.

Классы временных сессий:	**Однозадачные сессии**			**Линейные сессии (задачи исполняются одна за другой)**			**Многозадачные (задачи исполняются параллельно)**		
Набор данных	*Excite* 1999	*Excite* 2001	*Яндекс* 2005	*Excite* 1999	*Excite* 2001	*Яндекс* 2005	*Excite* 1999	*Excite* 2001	*Яндекс* 2005
% временных сессий, отнесенных к классу	87.56	86.87	86.40	11.32	11.94	12.53	1.12	1.19	1.07
Число задач во вр. сессии	1	1	1	2.20	2.21	2.19	2.34	2.35	2.30
Возобновление задачи новым запросом (%)	27.34	26.8	26.9	54.60	52.4	53.1	51.84	52.8	52.4
Число (несовпадающих) запросов в задаче	1.24	1.3	1.3	1.17	1.2	1.2	1.45	1.5	1.5
Число (несовпадающих) запросов во врем. сессии	1.24	1.3	1.3	2.56	2.6	2.6	3.40	3.5	3.4
Число транзакций одного запроса	1.56	1.7	1.6	1.52	1.6	1.5	1.87	2.0	2.0
Число транзакций в задаче	1.93	2.13	2.09	1.78	1.85	1.83	2.71	2.97	2.92
Продолжительность (мин) временной сессии:									
из 1 запроса	1.09	1.17	1.68						
из 2 запросов	4.37	4.45	5.60	5.84	5.72	6.89	10.62	11.52	13.55
из 3 запросов	7.35	7.36	8.97	9.27	8.87	11.04	11.16	11.79	14.30

Таблица 1. Результаты автоматического анализа трех логов

(*d'*) Качественное поведение характеристик в зависимости от типа временной сессии приведено на Рис. 4. На первый взгляд, наименее сложными в исполнении должны быть однозадачные сессии, сложность линейных сессий должна быть не ниже, и, наконец, параллельное исполнение задач наиболее сложно. Однако немонотонное поведение числовых характеристик на Рис. 4, говорит в пользу того, что субъективная сложность задач в однозадачных сессиях выше, чем у задач в линейных сессиях. Т.е. если текущая задача сложна для пользователя, то, скорее всего, он не начнет следующую.

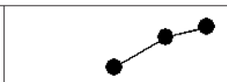
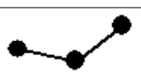
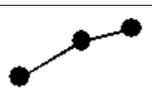
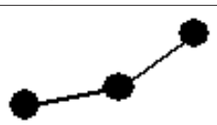
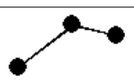
Характеристики, “монотонно” изменяющиеся по 3 классам временных сессий:		Характеристики, “немонотонно” изменяющиеся по 3 классам временных сессий:	
Число задач во временной сессии		Число несовпадающих запросов в задаче	
Число (уникальных) запросов в сессии		Число транзакций запроса	
Продолжительность сессии		Термов в запросе	
		Возобновление задачи новым запросом	

Рис. 4. Диаграммы изменения количественных характеристик по 3 классам временных сессий: левая из 3 точек соответствует значению характеристики на однозадачных сессиях, центральная – сессиям последовательного исполнения, правая – сессиям параллельного исполнения

012-Buzikashvili_H:_Layout 1 18.05.2008 14:40 Page 81

8. Ручная проверка

Была проведена выборочная проверка около 300 временных сессий, автоматически опознанных как многозадачные. Оказалось, что

(1) почти половина из них многозадачными не являются. Лексическое несходство запросов вызвано активным поиском, в котором пользователь, выполняя одну задачу, радикально меняет запросы и возвращается к запросам, похожим на ранее сделанные.

(1.1) среди временных сессий, автоматически опознанных как сессии ширины 2, настоящих многозадачных сессий чуть больше половины, и все они ширины 2;

(1.2) зато среди временных сессий, опознанных как более широкие, оказалось лишь несколько ширины 2, а все остальные были сессиями активного однозадачного поиска.

(2) характеристики действительно многозадачных сессий очень близки к характеристикам сессий, автоматически опознанных как многозадачные;

(3) среди действительно многозадачных сессий прерывающая задача допускает интерпретацию как связанная с прерванной (как ее ответвление или как поиск того, без чего продолжение прерванной затруднительно) редко (менее 20%). Обычно же прерванная и прерывающая задачи никак не связаны.

9. Заключение

Показано, что:

— многозадачность, понимаемая параллельное исполнение нескольких поисковых задач, – крайне редкая манера поиска;

— ширина многозадачных сессий равна двум, т.е. в любой момент есть не более одной начатой, но не завершенной поисковой задачи;

— возврату к прерванной задаче предшествует завершению ее прервавшей;

Таким образом, поисковое поведение пользователя подчиняется принципу минимума усилий: пользователь избегает многозадачности как более затратной, а прибегнув к ней, ограничивается исполнением только двух задач, причем наиболее экономным образом.

Параллельный многозадачный поиск, безусловно, факт, но столь редкий, что нет никакой необходимости в дальнейшем его изучении, по крайней мере, в контексте работ [11–14, 2,3]. Цель данной работы локальна – показать ошибочность мнения о массовости многозадачного поиска. Метод, использованный здесь, предназначен для решения именно этой задачи, которое никак не переносимо за ее рамки. Очень высокое число ошибок неотбраковки, не вызывающее затруднений в использованной процедуре, делает невозможным применение метода для автоматического решения другой, гораздо более важной проблемы – вычленения задач в последовательности всех запросов пользователя.

10. Благодарности

В заключение автор считает долгом выразить признательность Ю.Г. Зеленкову, Н.В. Пономаревой, А.В. Сокирко и А. Спинк за полезные обсуждения, а также анонимному рецензенту (в том числе, если он уже упомянут) – за ценные замечания по проблематике выявления задачных сессий.

Список литературы

1. Apostolico A.: String editing and longest common subsequences // Handbook of Formal Languages, vol, 2, Springer-Verlag, 1997, 361-398

2. Buzikashvili N. Automatic task detection in the Web logs and analysis of multitasking // 9th Int. Conf. on Asian Digital Libraries (ICADL 2006), LNCS 4312, Springer Verlag, 2006, 131-140.

3. Buzikashvili N., Ponomareva N.V. Multitasking search on the Web // 7th Int. Conf. on Cognitive Modeling (ICCM 2006). Edizioni Goliardiche, 2006, 357-358.

4. Buzikashvili N. Original and normalized Web log metrics as functions of controllable variables of log study // 5th Latin American Web Conference (LA-WEB 2007), IEEE CS, 2007, 3-12.

5. Kaki M. fKWIC: Frequency-based keyword-in-context index for filtering Web search results // JASIST, 52(12), 2006, 1606-1615

6. Kushleyeva Y., Salvuci D, Lee FJ. Deciding when to switch tasks in time-critical multitasking // Cognitive Systems Research, 6, 2005, 41-49.

012-Buzikashvili_H:_Layout 1 18.05.2008 14:40 Page 82

7. Monsell S.: Task switching // Trends in Cognitive Sciences, 7(3), 2003, 134-140

8. Montgomery A., Faloutsos C. Identifying Web browsing trends and patterns // Computer 34 (7), 2001, 94-95.

9. Rubinstein, J., Meyer, D., Evans, J. Executive control of cognitive processes in task switching // J. Exp. Psycol. Hum. Percept. Performance, 27, 2001, 763-797.

10. Silverstein, C., Henzinger, M., Marais, H., Moricz, M.: Analysis of a very large web search engine query log // SIGIR Forum, 33 (1), 1999,: 6–12.

11. Spink A., Jansen B.J., Pedersen J. Multitasking Web search on Alta Vista // Int. Conf. on Information Technology: Coding and Computing (ITCC'04). IEEE CP. 2004, 309-313.

12. Spink A., Jansen B.J., Park M., Pedersen J. Multitasking during Web search sessions // Information Processing & Management, 42(1), 2006, 264-275.

13. Spink A, Koshman S, Park M, Field C, Jansen B.J. Multitasking Web Search on Vivisimo.com // Int. Conf. on Information Technology: Coding and Computing (ITCC'05), Vol. II, IEEE CS, 2005, 486-490.

14. Spink A., Cole C., Waller M. Multitasking behavior. // Chapter 4. Annual Review of Information Science and Technology (ARIST) vol. 42, 2008, Medford, NJ: Information Today, Inc. (to appear).

013-Vasiliev_V:_Layout 1 18.05.2008 14:26 Page 83

КОМПЛЕКСНАЯ ТЕХНОЛОГИЯ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕКСТОВ

COMPLEX TECHNOLOGY OF AUTOMATIC TEXT CLASSIFICATION

Васильев В.Г. *(vvg_2000@mail.ru)*
Институт прикладной информатики РАН

В докладе рассматриваются проблемы, которые возникают при построении прикладных систем автоматической классификации текстов. Приводится описание основных элементов комплексной технологии классификации текстов, обеспечивающей полный цикл обработки и анализа данных, начиная с очистки и выделения текста из документов, заканчивая анализом результатов классификации. Особое внимание уделяется вопросам построения комбинированных решающих правил для выполнения иерархической классификации текстов.

Введение

Потребность в автоматизации различных задач, связанных с обработкой и анализом текстовых данных на естественном языке, испытывают как рядовые пользователи средств вычислительной техники, так и крупные государственные и частные организации. В области автоматизированной обработки текстов уже сложился ряд относительно самостоятельных направлений (задач): извлечение объектов и признаков, реферирование, классификация, кластерный анализ, интеллектуальный поиск, фактографический анализ, пространственный (географический) анализ. В настоящей работе указанные базовые задачи анализа текстов рассматриваются не независимо друг от друга, а как элементы комплексной технологии автоматической классификации текстовых данных, обеспечивающей эффективную обработку информации и представление результатов анализа для конечных пользователей (см. рис. 1).

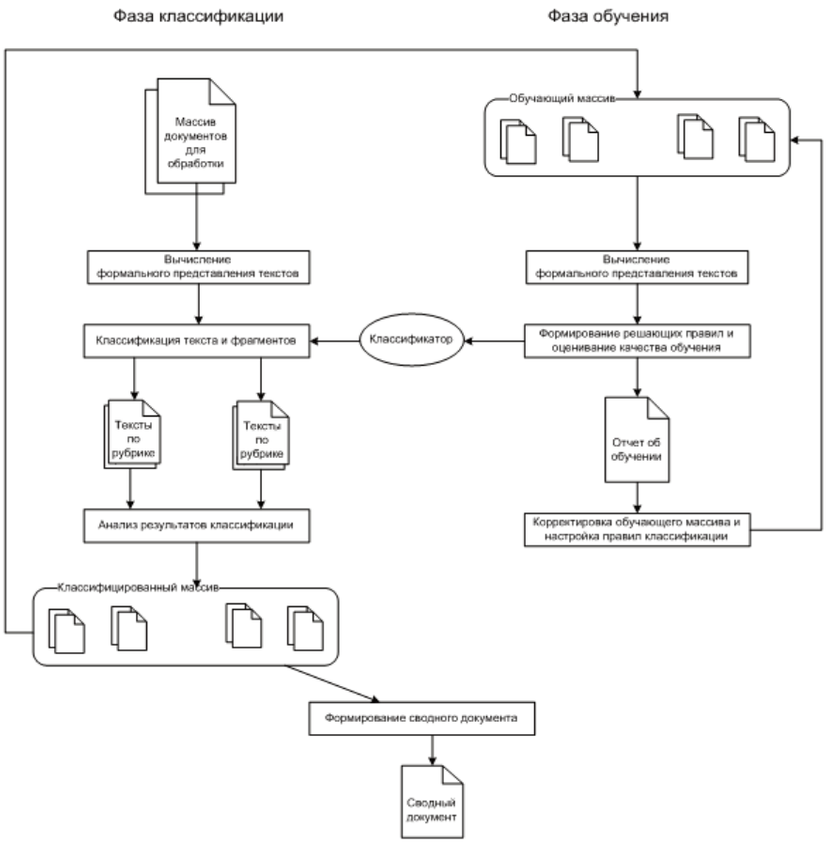


Рис. 1. Общая схема комплексной технологии классификации текстов

013-Vasiliev_V:_Layout 1 18.05.2008 14:26 Page 84

Краткое описание задач, решаемых в рамках приведенных на рис. 1 функциональных блоков приводится в табл. 1.

№	Функциональный блок	Задачи функционального блока
1	Вычисление формального представления текстов	1. Получение текста документа (идентификация формата, языка, кодировки документа, очистка текста от элементов оформления, разбиение на составные части). 2. Лингвистический анализ (графематический, морфологический и постморфологический анализ, выделение словосочетаний). 3. Формирование векторного (матричного) представления текстов.
2	Классификация текстов и фрагментов	1. Предварительная обработка текстов (отображение словаря признаков документа в пространство признаков классификатора; оценка адекватности и возможности классификации текста с помощью данного классификатора). 2. Классификация текста и выделение значимых фрагментов в нем (выделение кодов рубрик с помощью регулярных выражений; применение логических правил, построенных экспертами, и статистических решающих правил; корректировка результатов классификации с учетом иерархической структуры рубрик).
3	Анализ результатов классификации	1. Выявление "почти" дубликатов документов. 2. Выявление основных тем документов в рубриках..
4	Формирование сводного документа	1. Формирование сводного документа (упорядочение документов по их релевантности рубрике; упорядочение фрагментов с учетом их близости друг к другу; построение объединенного документа).
5	Формирование решающих правил и оценивание качества обучения	1. Формирование обучающих и тестовых множеств для рубрик (построение разбиения обучающего массива на блоки; анализ взаимосязей и пересечений отдельных рубрик; формирование множеств отрицательных и положительных примеров). 2. Оценивание параметров базовых моделей рубрик (вычисление весов признаков; снижение размерности; оценивание параметров моделей; формирование решающих правил; оценка качества обучения). 3. Построение комбинированных решающих правил для отдельных рубрик и классификатора в целом. 4. Формирование отчета о результатах обучения (описание решающих правил, описание терминологии рубрик, рекомендации по корректировке примеров документов, описание взаимосвязей рубрик).
6	Корректировка обучающего массива и настройка правил классификации	1. Корректировка обучающих примеров для рубрик путем анализа добавленных и пропущенных документов в рубриках, значимых фрагментов, взаимосвязей рубрик. 2. Настройка правил классификации для отдельных рубрик (явное задание предпочтительных статистических моделей, задание необходимых, достаточных и исключающих логических правил на специальном языке).

Таблица 1. Описание функциональных блоков комплексной технологии

Описанная выше технология полностью реализована как в виде отдельного пакета «Text Classification Toolbox« для системы Matlab [9], так и в виде элемента ряда заказных информационно-аналитических систем.

При ее построении особое внимание было уделено учету особенностей реальных массивов текстов, которые часто оказывает негативное влияние на итоговое качество классификации текстов. В табл. 2 приведено описание наиболее типичных из них (отмечаются во многих работах и наблюдались автором на практике), а также способы их учета в рамках предлагаемой технологии.

013-Vasiliev_V:_Layout 1 18.05.2008 14:26 Page 85

№	**Название особенностей и недостатков (ссылки на работы)**	**Проблемы и средства их решения в рамках комплексной технологии**
1	Наличие недостаточного количества или отсутствие обучающих примеров для ряда рубрик ([1], [3], [5], [6]).	Проблемы: невозможность построения правил классификации для большинства методов, основанных на обучении по примерам; низкая надежность оценки качества обучения. Решение: - поддержка совместного использования трех типов решающих правил для рубрик: статистических (обучаемых на примерах документов), логических (задаются экспертами на специальном информационно-поисковом языке), шаблонных (задаются экспертами в виде регулярных выражений).
2	Наличие ошибок и непоследовательность формирования эталонного распределения текстов по рубрикам ([1], [2], [5], [6]). Односторонность примеров в обучающем массиве текстов ([1]). Наличие рубрик с одинаковым содержанием, но с разными названиями.	Проблемы: формирование ошибочных правил классификации; результаты оценки качества обучения оказываются некорректными. Решение: - автоматическое выполнение при обучении оценки качества классификации и ошибок в эталонном распределении документов по рубрикам; - формирование обучающих примеров для отдельных рубрик с учетом оценки степени тематической близости рубрик друг к другу; - реализация оригинальной интерактивной процедуры обучения классификатора, обеспечивающей корректировку обучающих примеров и задание логических правил.
3	Несоответствие тематики и характера обучающего массива текстов тематике и характеру классифицируемых текстов ([5], [6]).	Проблемы: результаты классификации текстов могут быть неопределенными и зависеть от незначительных случайных факторов; результаты оценки качества обучения являются завышенными. Решение: - выполнение оценки качества классификации в процессе обучения; - обеспечение переобучения в процессе обработки новой информации; - использование дополнительных словарей квазисинонимов для повышения полноты классификации.
4	Многоуровневый (иерархический) характер классификаторов ([1], [5]). Совместное использование нескольких оснований (принципов) разделения данных на классы (тема и тональность). Неоднородность характера текстов в рубриках.	Проблемы: сложность построения эффективных процедур классификации, основанных на использовании одной модели или метода для всех рубрик и уровней классификатора. Решение: - поддержка нескольких типов признаков (лексических, грамматических, синтаксических); - использование оригинального комбинированного иерархического метода классификации, обеспечивающего подбор для каждой рубрики наиболее подходящих для нее моделей и методов классификации [4].
5	Одновременная оценка текстов с использованием нескольких классификаторов (например, страна, организация, тема, жанр) [7].	Проблемы: сложность подбора примеров документов и обучения объединенного классификатора, получаемого путем декартового произведения рубрикаторов для отдельных фасетов; сложность или невозможность независимого обучения классификаторов из-за наличия контекстных связей между рубриками классификаторов. Решение: - поддержка режима фасетной классификации, который обеспечивает независимое обучение отдельных фасетов и комбинирование их работы с использованием специальных логических правил.

013-Vasiliev_V:_Layout 1 18.05.2008 14:26 Page 86

№	Название особенностей и недостатков (ссылки на работы)	Проблемы и средства их решения в рамках комплексной технологии
6	Политематический и составной характер документов ([1]). Наличие служебных элементов и посторонних блоков текста, не относящихся к основной тематике документа [1]. Наличие в обучающем массиве аномальных документов (пустых, в неизвестных кодировках и т.п.) ([2]).	Проблемы: сложность автоматического формирования решающих правил для рубрик из-за негативного влияния посторонней информации; снижение качества классификации из-за наложения нескольких рубрик друг на друга; сложность интерпретации результатов классификации из-за неопределенности расположения в тексте информации, релевантной рубрике. Решение: - реализация полного комплекса средств для идентификации форматов, языков и кодировок документов; - реализация оригинальных алгоритмов для очистки текста документов от элементов оформления, основанных на оценке распределения плотности текста на странице; - реализация оригинальных алгоритмов для исключения из текстов вспомогательной информации, основанных на сопоставлении лексического состава отдельных предложений; - реализация робастных вариантов алгоритмов оценивания параметров моделей и методов классификации [4, 11]; - реализация эффективного выделения значимых фрагментов в текстах путем использования оригинальной иерархической модели представления текстов (основывается на представлении текста множеством векторов признаков, соответствующих элементам специального иерархически упорядоченного перекрывающегося множества фрагментов текста) [10].
7	Наличие повторяющейся информации во входном потоке текстов [8]. Большие объемы и неоднородность входного потока текстов ([3]).	Проблемы: сложность просмотра и анализа результатов классификации. Решение: - реализация оригинального алгоритма упорядочения документов в рубриках, который учитывает не только релевантность документов рубрике, но и их тематическую близость друг к другу; - реализация оригинального алгоритма для выявления "почти дубликатов" документов, основанного на использовании методов иерархического кластерного анализа и иерархической модели представления текстов; - реализация оригинальных алгоритмов кластерного анализа результатов классификации, основанных на использовании модели смеси вероятностных анализаторов главных компонент и обеспечивающих эффективное формирование графических карт массивов текстов [11].

Таблица 2. Типичные особенности и недостатки реальных массивов текстов

Полное описание всех элементов комплексной технологии сложно провести в рамках одной работы. По этой причине остановимся более подробно на рассмотрении одного из ее базовых элементов – методе комбинированной иерархической классификации.

Описание метода комбинированной иерархической классификации

Разнородный характер текстов в рубриках и использование различных оснований классификации приводит к сложности выделения метода классификации, который был бы эффективным во всех случаях. Возможным решением является совместное использование сразу нескольких методов классификации текстов [4].

В общем виде работу комбинированного иерархического классификатора можно представить следующим образом. На вход поступает вектор весов информационных признаков анализируемого текста или фрагмента, а на выходе формируются два вектора $c = (c_1,...,c_k)$ и $w = (w_1,...,w_k)$, где общее число рубрик в классификаторе, $c_j \in \{0,1\}$ и $w_j \in [0,1]$ – признак и степень принадлежности к рубрике ω_j, соответственно.

013-Vasiliev_V:_Layout 1 18.05.2008 14:26 Page 87

Решающие правила для отнесения текстов к рубрикам получаются путем комбинирования результатов работы сразу нескольких базовых методов классификации с помощью метаклассификаторов более высокого уровня. Всего выделяется три уровня: уровень базовых классификаторов, уровень комбинированных классификаторов, уровень иерархического классификатора.

На первом уровне для каждой рубрики ω_j, $j = 1,...,k$ производится построение бинарных решающих с помощью следующих базовых методов [1, 2, 3]:

- методы вероятностной классификации, основанной на представлении рубрик в виде смеси распределений Бернулли (BERN), фон Мизеса-Фишера (VMF), полиномиального (MNS) и анализаторов главных компонент (PPCA);

- методы классификации на основе вычислений расстояний: классификаторы k – ближайших соседей (KNN), машин опорных векторов (SVM), Роччио (ROC);

- методы классификации на основе правил: деревья решений (TREE), логические правила на специальном языке.

Все приведенные методы, за исключением логических правил, основаны на обучении на примерах. При этом при обучении для каждого метода реализуется специализированная процедура обработки данных, которая включает проверку достаточности размера обучающей выборки, снижение размерности путем селекции и трансформации признаков, оценивание параметров, построение решающих правил, оценка качества обучения (особенности реализации и параметры по умолчанию для базовых методов показаны в табл. 3).

Метод	**Мин. и макс. размер обуч. множества**	**Веса признаков**	**Снижение размерности**	**Оценка параметров**	**Решающее правило**
BERN	2 - 50000	Бинарные [1]	селекция по частоте документов [1]	байесовское оценивание	байесовское правило с откл. вер. уровня 65% [2]
VMF	5 - 50000	TF-IDF [1]	селекция по частоте документов [1]	Оригинальный робастный алгоритм	байесовское правило с откл. вер. уровня 95% [2]
MNS	2 - 50000	TF-IDF [1]	селекция по частоте документов [1]	Оригинальный робастный алгоритм	байесовское правило с откл. вер. уровня 80% [2]
PPCA	10 - 50000	TF-IDF [1]	селекция по частоте документов, последовательный метод LSI [11]	оригинальный робастный алгоритм [11], размерность пространства факторов – 5.	байесовское правило с откл. вер. уровня 60% [2]
KNN	5 - 50000	TF-IDF [1]	селекция по частоте документов [1]	число соседей – 5, максимальное число эталонов – 250, оригинальный алгоритм отбора эталонов на основе кластерного анализа	байесовское правило с откл. вер. уровня 60% [2]
SVM	5 - 50000	TF-IG [1]	селекция по частоте документов [1]	линейная ядерная функция [2]	линейная решающая функция
ROC	2 - 50000	TF-IDF [1]	селекция по частоте документов [1]	стандартный алгоритм [1]	линейная решающая функция
TREE	10 - 50000	Бинарные - IDF [1]	селекция по частоте документов, метод фильтрации признаков Information Gain [1]	оригинальный алгоритм отбора эталонов на основе кластерного анализа, стандартный алгоритм из пакета matlab	логическое решающее правило

Таблица 3. Особенности реализации базовых методов классификации

013-Vasiliev_V:_Layout 1 18.05.2008 14:26 Page 88

Логические правила строятся вручную для уточнения и дополнения статистических решающих правил, а также построения правил для рубрик без примеров документов. Они разбиваются на три типа: достаточные (справедливость достаточна для отнесения текста к рубрике), необходимые (для отнесения текста к рубрике необходима справедливость данного правила и построенного статистического правила), отрицательные (при его справедливости текст не относится к рубрике).

Реализованный язык задания логических правил обеспечивает поиск отдельных слов (с учетом и без учета морфологии, с учетом ошибок, с заданными морфологическими и семантическими характеристиками), задание логических условий, задание условий на расстояние между выражениями в тексте.

На втором уровне для каждой рубрики ω_j, $j = 1,...,k$ осуществляется построение отдельного комбинированного классификатора на основе бинарных классификаторов первого уровня $C_{j1},...,C_{jL}$, построенных для данной рубрики, где L – число различных методов классификации. Для этих целей реализована поддержка нескольких методов, которые условно можно разбить на три группы [2, 4]:

- методы, основанные на построении фиксированного решающего правила, которое не зависит от качества работы отдельных классификаторов (например, правило произведения апостериорных вероятностей, правило суммирования апостериорных вероятностей, правило большинства голосов, правила минимума апостериорных вероятностей классов).

- методы, основанные на построении комбинированного правила классификации, учитывающего оценки качества работы классификаторов первого уровня (например, байесовский метод и метод наилучшего классификатора).

- методы, основанные на использовании статистического моделирования (например, boosting и bagging).

Экспериментальная оценка приведенных групп правил показала, что использование методов первой группы не приводит к улучшению качества классификации, но при этом время работы и требования к памяти заметно возрастают из-за необходимости одновременного использования для каждой рубрики нескольких алгоритмов при классификации. Методы третьей группы требуют выполнения чрезвычайно ресурсоемкого моделирования, которое не позволяет проводить обучение классификаторов за разумное время на практике.

В настоящей работе в качестве основного был выбран метод наилучшего классификатора, в рамках которого комбинированный классификатор для рубрики ω_j, $j = 1,...,k$ получается путем выбора из $C_{j1},...,C_{jL}$, классификатора, обеспечивающего наибольшее значение F-меры на тестовом множестве.

Такой подход обладает следующими преимуществами:

- при классификации для каждой рубрики требуется хранить в памяти только одно решающее правило,

- результаты оценки качества, проводимой для выбора наилучшего метода, могут использоваться для корректировки состава обучающих примеров.

На третьем уровне осуществляется построение иерархического классификатора, объединяющего результаты работы бинарных классификаторов второго уровня, таким образом, чтобы обеспечить отнесение текстов одновременно к нескольким рубрикам с учетом их иерархической структуры. Его работа сводится к выполнению серии процедур, которые, например, осуществляют проверку при отнесении документа к рубрике, что он относится к родительской рубрике, производят проверку различных аномальных случаев.

Для оценки эффективности разработанной технологии были проведены эксперименты с различными массивами текстов и рубрикаторами. К сожалению, большинство из рассмотренных массивов не являются общедоступными, что затрудняет публикацию информации по ним. По этой причине в качестве примера приведем результаты экспериментов только с массивом «Reuters-21578», который широко используется в различных работах по автоматизированной обработке текстов [1].

Для проведения экспериментов использовался пакет «Text Classification Toolbox», реализующий описанный выше комбинированный иерархический алгоритм классификации. При обучении классификатора в рамках данного пакета автоматически производится формирование отчета с оценками качества работы базовых методов и классификатора в целом с использованием метода 5-шаговой кросс проверки. Для упрощения проведения экспериментов было решено не использовать эталонное разбиение на обучающую и тестовую выборку, которое имеется в массиве «Reuters-21578». Это может приводить к несколько отличным от других исследователей абсолютным значениям показателей качества классификации, но это не является критичным, так как в данном случае для иллюстрации наибольшее значение имеют относительные значения показателей. Все базовые алгоритмы использовались с приведенными выше значениями параметров.

В табл. 4 приводятся оценки качества обучения 13 из 142 рубрик данного массива с помощью 8 методов классификации.

013-Vasiliev_V:_Layout 1 18.05.2008 14:26 Page 89

Рубрика (размер)	**TREE**	**PPCA**	**MNS**	**KNN**	**BERN**	**VMF**	**ROC**	**SVM**
acq (2261)	85%	95%	40%	54%	92%	95%	3%	98%
alum (59)	91%	85%	59%	83%	73%	89%	54%	94%
dmk (15)	76%	88%	88%	92%	64%	97%	92%	88%
housing (18)	85%	84%	84%	94%	81%	91%	88%	88%
l-cattle (10)	-	-	36%	95%	67%	90%	88%	29%
meal-feed (51)	97%	93%	63%	80%	81%	94%	77%	94%
palm-oil (42)	85%	98%	85%	91%	82%	94%	91%	94%
propane (6)	-	-	-	-	55%	-	80%	-
rapeseed (35)	72%	86%	76%	90%	94%	91%	75%	90%
sfr (3)	-	-	100%	-	80%	-	80%	-
soy-oil (26)	22%	36%	7%	20%	51%	40%	33%	26%
strategic-metal (39)	75%	80%	32%	61%	60%	77%	51%	73%
zinc (48)	74%	85%	60%	70%	74%	91%	78%	81%

Таблица 4. Качество обучения отдельных рубрик с помощью различных методов

Необходимо отметить, что у ряда рубрик в массиве «Reuters-21578« отсутствуют примеры документов или их количество является недостаточным для оценивания параметров моделей данных применяемых в отдельных методах классификации. Такие случаи отмечены в таблице прочерками.

Из табл. 4 видно, что для каждого метода классификации существуют рубрики, на которых он оказывается значительно предпочтительнее других методов.

В табл. 5 приводятся усредненные оценки качества обучения комбинированного классификатора для следующих случаев: используется только один базовый метод классификации, используются все базовые методы без задания и с заданием экспертных логических правил. Оценки коэффициентов точности, полноты и F-меры вычислялись с использованием микро-усреднения [1, 3].

Метод классификации	**Точность (дов. инт.)**	**Полнота (дов. инт.)**	**F-мера**	**Процент обученных рубрик**
TREE	81% (80%, 82%)	87% (86%, 87%)	84%	50%
PPCA	94% (93%, 94%)	95% (95%, 95%)	94%	50%
MNS	66% (65%, 67%)	60% (60%, 61%)	63%	65%
KNN	76% (76%, 77%)	73% (72%, 74%)	75%	56%
BERN	81% (80%, 82%)	87% (86%, 87%)	84%	70%
VMF	92% (91%, 92%)	93% (92%, 93%)	92%	56%
ROC	41% (40%, 42%)	36% (35%, 36%)	38%	70%
SVM	95% (95%, 95%)	96% (95%, 96%)	95%	56%
Комбинированный метод (без логических правил)	97% (98%, 99%)	97% (98%, 99%)	97%	70%
Комбинированный метод (с логическими правилами)	98% (98%, 99%)	99% (98%, 99%)	99%	100%

Таблица 5. Оценка качества обучения с использованием микро-усреднения

Из табл. 5 видно, что среди базовых методов наилучшие результаты показал метод SVM, что в целом согласуется результатам экспериментов других авторов с данным массивом текстов [1]. Использование комбинированного метода позволяет дополнительно повысить качество классификации. При этом наилучшие результаты достигаются в том случае, когда дополнительно для рубрик вручную задаются логические правила на специальном языке, которые корректируют ошибки статистической классификации. Необходимо отметить, что в результате задания таких правил также удается построить решающие правила для рубрик, у которых отсутствуют обучающие примеры или их меньше 2.

013-Vasiliev_V:_Layout 1 18.05.2008 14:26 Page 90

Влияние необученных рубрик и рубрик небольшого размера на показатели качества обучения более отчетливо проявляется при использовании макро-усреднения. В данном случае, обобщенные показатели вычисляются как арифметическое среднее показателей для отдельных рубрик. В результате, значение F-меры для метода SVM (наилучший метод при микроусреднении) становится равным 49%, а для комбинированного метода – 63% (без задания экспертных логических правил) и 85% (с заданием экспертных логических правил).

Значительное отличие микро и макро усреднения связано с тем, что в массиве имеется несколько больших рубрик, на которых достигаются высокие показатели качества классификации (например, для рубрики «acq» F-мера равна 99%), и большое количество пустых и маленьких рубрик, на которых F-мера принимает значения близкие к нулю. Повышенные значения качества работы комбинированного алгоритма с экспертными логическими правилами объясняются тем, что за счет подбора данных правил достаточно легко обеспечить точность и полноту классификации близкую к 100% для рубрик состоящих всего из нескольких документов.

Заключение

Таким образом, в настоящей работе рассмотрены базовые элементы комплексной технологии классификации текстов, которая реализована в виде пакета для системы Matlab. Отличительной особенностью данной технологии является ориентация на совместное применение экспертных и статистических методов классификации текстов, а также интегрированное использование всего множества процедур автоматизированной обработки текстов, начиная с очистки текстов от посторонней информации, заканчивая интерпретацией результатов. В частности, приведенные эксперименты с массивом «Reuters-21578« показали перспективность подхода, основанного на комбинировании методов классификации текстов.

К перспективным задачам можно отнести следующие: разработка эффективных методов обучения фасетных классификаторов при наличии выборок ограниченного объема и большом количестве фасетов; совершенствование процедур комбинирования результатов работы экспертных и статистических процедур классификации за счет задания соответствующих правил на специальном языке; разработка эффективных процедур и методик автоматизированного формирования сводных документов по результатам автоматической классификации.

Работа выполнена при поддержке гранта Президента Российской Федерации для государственной поддержки молодых российских ученых МК-12.2008.10.

Список литературы

1. Sebastiani F. Machine learning in automated text categorization // ACM Computing Surveys, 34(1), 2002. – pp. 1-47.

2. Webb R.A. Statistical Pattern Recognition. Second Edition. // John Wiley & Sons Ltd., England, 2002. – 515 p.

3. Baldi P., Frasconi P., Smyth P. Modeling the Internet and theWeb. Probabilistic Methods and Algorithms // JohnWiley & Sons Ltd, 2003. – 306 p.

4. Кривенко М.П., Васильев В.Г. Проблемы разработки и внедрения технологий извлечения информации // Системы высокой доступности 3-4, т.2. – М.: Радиотехника, 2006. – с. 6-21.

5. Агеев М.С., Добров Б.В., Лукашевич Н.В. Поддержка системы автоматического рубрицирования для сложных задач классификации текстов // Труды 6-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL2004, Пущино, Россия, 2004. – 10 с.

6. Dumais S.T., Lewis D.D. & Sebastiani F., Report on the Workshop on Operational Text Classification Systems // SIGIR-02, Tampere, Finland. – 4 p. (http://www.sigir.org/forum/F2002/sebastiani.pdf).

7. Браславский П.И. Использование стилистических параметров документа при поиске информации в Internet // Доклады VI рабочего совещания по электронным публикациям – EL-PUB–2001. Новосибирск: ИВТ СО РАН, 2001. (http://www.ict.nsc.ru/ws/elpub2001/1812).

8. Yang H., Callan. J. Near-Duplicate Detection for eRulemaking // Proceedings of the 5th National Conference on Digital Government Research (DG.O2005), Atlanta, GA, USA, 15-18 May 2005.

9. Васильев В.Г., Кривенко М.П., Ефременкова М.В. Библиотека процедур классификации тексто-вых данных // Редакция «ОПиПМ» Обозрение приклад-ной и промышленной математики. Выпуск 1, том 13. – М., 2006. – с. 743.

10. Васильев В.Г. Автоматическое выделение значимых фрагментов в текстах // Редакция «ОПиПМ» Обозрение прикладной и промышленной математики. Выпуск 3, том 14. – М., 2007. – с. 518.

11. Кривенко М.П., Васильев В.Г. Кластерный анализ массивов текстовых данных // Препринт. – М.: ИПИ РАН, 2004. – 190 с.

014-Voskresenskiy_AL:_Layout 1 18.05.2008 14:30 Page 91

СОПОСТАВИТЕЛЬНОЕ ЛЕКСИКОГРАФИЧЕСКОЕ ОПИСАНИЕ СЛОВ РУССКОГО ЯЗЫКА И ЖЕСТОВ ЯЗЫКА ГЛУХИХ РОССИИ В СЛОВАРЕ RUSLED

COMPARATIVE LEXICOGRAPHIC DESCRIPTION OF RUSSIAN WORDS AND GESTURES OF RUSSIAN SIGN LANGUAGE IN RUSLED DICTIONARY

Воскресенский А.Л. *(AVoskresenskij@college.mesi.ru)*
Колледж МЭСИ, Москва

Представляется лексикографическое описание жестов глухих России в сопоставлении со словами русского языка, имеющими те же или близкие значения, в мультимедийном толковом словаре RuSLED, предназначенном для изучения особенностей употребления жестов и слов русского языка.

Введение

Двуязычный словарь является одним из мест встречи двух культур, языки которых представлены в словаре. Он должен не только дать справку об употреблении речевых единиц (в данном случае слов и жестов) соответствующих языков, но и представить семантическую структуру отдельных речевых единиц с учётом оттенков и переходов в их значениях и употреблении. При этом справочная информация словаря должна быть понятна пользователю с учетом существующих культурных различий.

Различия между социальными группами слышащих и глухих достаточно велики [1]. Учитывая это, а также то, что между словами русского языка и жестами языка глухих России во многих случаях нет однозначного соответствия [2], нужно признать, что существующие словари жестового языка (см., например, [3, 4]) не дают достаточно ясного (для человека, не владеющего языком жестов) описания особенностей применения жестов. Кроме того, для многих глухих было бы полезно иметь толковый словарь русского языка, в котором слова дополнительно пояснялись бы жестами.

Нужно признать, что составители словаря RuSLED не являются специалистами в лексикографии, и данный словарь является их первым опытом в данной области. Но это имеет и положительные черты. Например, составители словарей жестов обычно дают варианты жестов без пояснений, являются ли эти варианты диалектными формами или несут различные оттенки значений. Для них это очевидно, поэтому необходимые для новичка в данной области пояснения опускаются. С другой стороны, при описании слов русского языка, в связи с многообразием предметных областей, в которых фигурируют слова-омонимы, даже опытные специалисты иногда ошибаются [5], особенно в случаях, когда решение кажется очевидным. В этом случае действует механизм, сходный с «ложными друзьями переводчика» при переводе с одного языка на другой [6].

Для разработчиков словаря RuSLED все в данной области деятельности является новым, поэтому даже в очевидных для специалистов случаях у них возникают «детские» вопросы *«почему?», «чем это объясняется?»* и т.п. Может быть именно поэтому словарь RuSLED не имеет известных аналогов, в которых такое внимание уделялось бы толкованиям слов и жестов. При всех своих недостатках словарь вызвал интерес у представителей общины глухих Москвы и даже такие высокие оценки, как «это то, что нужно».

Описание словаря RuSLED

Словарь русского жестового языка включает в себя функции толкового словаря, как для введенного слова, так и для его жестового представления. На вход словаря подается произвольная форма слова, а на выходе демонстрируются варианты жестового толкования данного слова.

014-Voskresenskiy_AL:_Layout 1 18.05.2008 14:30 Page 92

Для нормализации входной словоформы (получения лексемы) разработан морфологический анализатор, в основу которого положен морфологический анализатор [7], использованный ранее в разработках TULIPS, TULIPS-2 и др. В качестве основного источника допустимых словоизменений (также как в [7] и в большинстве отечественных систем морфологического анализа) используется словарь А.А. Зализняка [8].

Морфологический анализатор включен в словарь как элемент будущей системы перевода текста в жесты.

Отличием разрабатываемого словаря является то, что для каждого семантического значения лексемы (и жеста) используется отдельный вход словаря – отдельная запись в таблице базы данных. Это значительно удобнее для пользователя, является очевидным решением для электронных толковых словарей и рекомендуется лексикографами [9].

В соответствии с этим изменены по сравнению с [7] как структура базы данных, так и запросы на выборку данных, реализованные на языке SQL. Использование SQL облегчит предполагаемую в дальнейшем реализацию мультимедийного толкового словаря русского жестового языка в виде Web-сервиса.

Форма доступа к данным демонстрационной версии словаря, выполненной на СУБД MS Access, приведена на рис. 1.

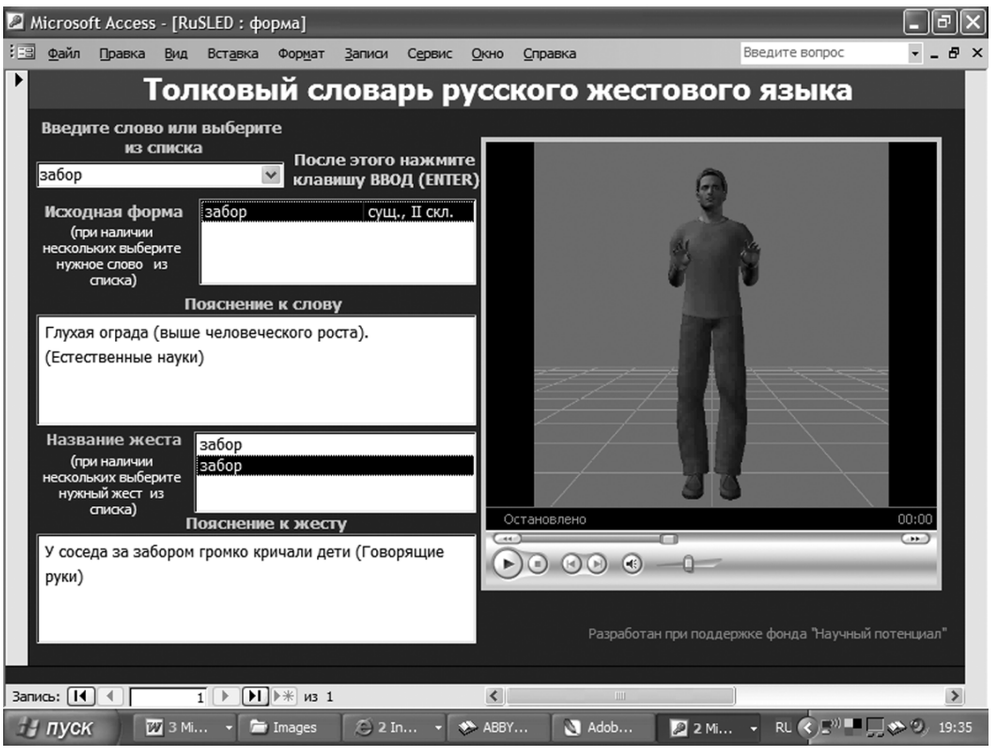


Рис. 1. Словарь RuSLED. Форма доступа к данным.

Поле «Введите слово» позволяет вводить произвольные словоформы или выбирать из списка лексемы, имеющиеся в словаре. В список «Исходная форма» выводится соответствующее основе значение лексемы или несколько значений, если по результатам морфологического анализа выбрано несколько записей, рис. 2, на котором приведен случай ввода пользователем словоформы «бора». Она соответствует именительному падежу слова «бора», обозначающего сильный ветер в приморских районах, где невысокие горы подступают непосредственно к побережью, а также родительному падежу слова «бор», имеющему несколько значений: еловый или

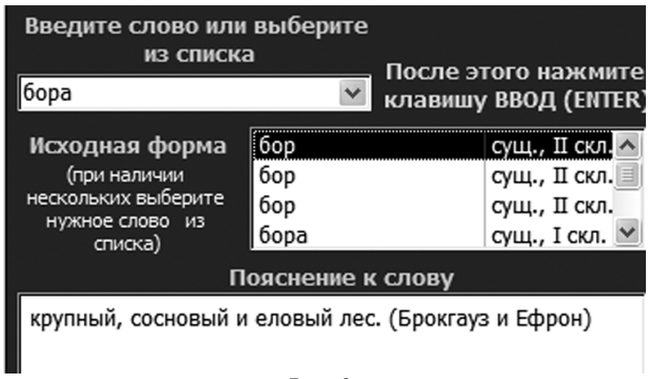


Рис. 2.

014-Voskresenskiy_AL:_Layout 1 18.05.2008 14:30 Page 93

сосновый лес, химический элемент, инструмент стоматолога и т.д. При выборе пользователем элемента списка «Исходная форма» в поле «Пояснение к слову» выводится толкование выбранной лексемы.

На рис. 3, 4 приведен случай ввода пользователем словоформы «стекла». Морфологический анализатор в поле «Исходная форма» выводит существительное «стекло» и глагол «стечь». Пример показывает, что морфологический анализатор словаря учитывает чередование согласных в различных словоформах лексемы. Учитываются пропуски букв («беглые гласные»), добавление возвратных частиц («-ся», «-сь»).

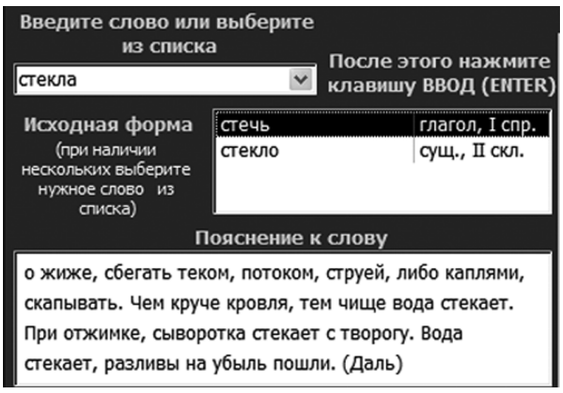


Рис. 3.

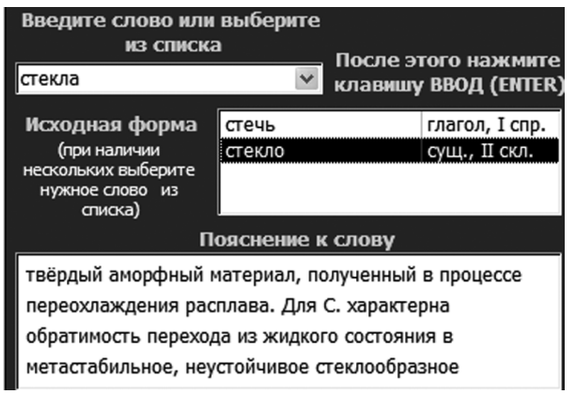


Рис. 4.

При выборе пользователем нужной лексемы в поле «Название жеста» выводится наименование жеста (как правило, совпадающее с лексемой) или (если данной лексеме соответствуют несколько жестов) список наименований (рис. 1). При выборе пользователем нужного жеста его изображение выводится в окне плеера, а в поле «Пояснение к жесту» выводится поясняющий текст. В ходе разработки осуществлена оцифровка [4], что позволяет использовать фрагменты этого курса в словаре RuSLED (рис. 4, 5). Постепенно осуществляется замена видеофрагментов анимированными изображениями с целью перехода к компоновке жестовых высказываний из комбинаций жестов с использованием единого демонстратора жестов – виртуального персонажа.

На рис. 5 и 6 представлены результаты выбора пользователем различных значений слова «лук». Для каждого из значений этого слова выдается только то значение жеста, семантика которого отвечает выбранной лексеме. Подробнее этот механизм будет рассмотрен ниже.

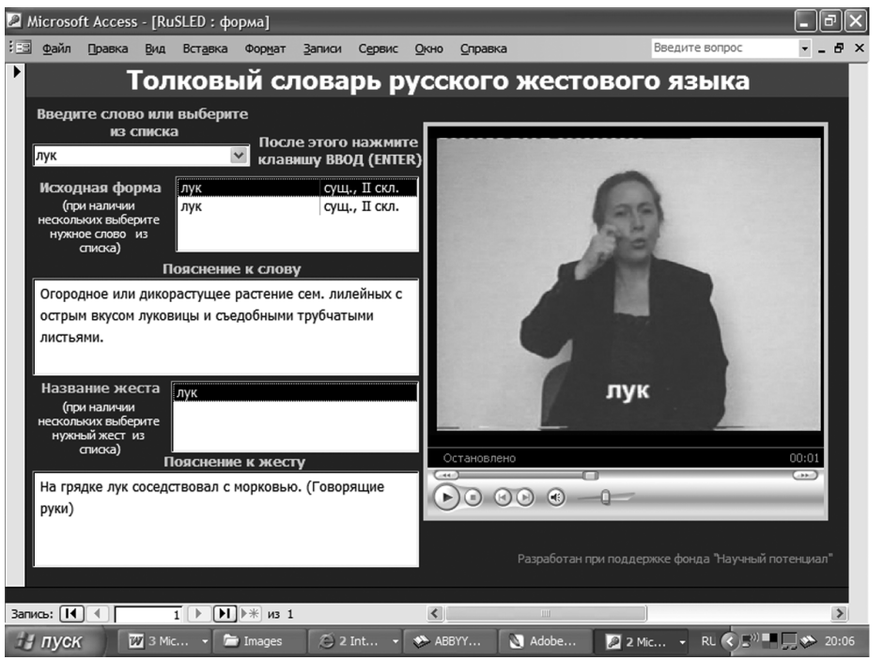


Рис. 5. Жест «лук» (растение).

014-Voskresenskiy_AL:_Layout 1 18.05.2008 14:30 Page 94

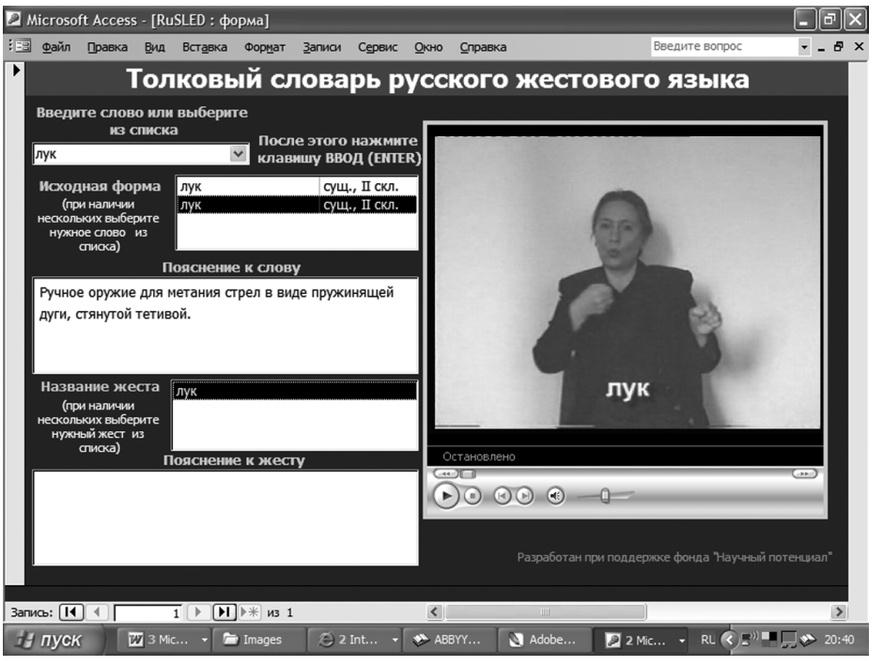


Рис. 6. Жест «лук» (оружие).

Источник словника словаря RuSLED

Одной из важнейших (и сложных) задач является определение размера и состава словника словаря. Поскольку предполагается, что данный словарь должен служить пособием при изучении русского жестового языка, словник (и *жестовник* ?) должны обеспечивать хотя бы минимальные требования коммуникации.

Первоначально словарь предполагалось создавать на основе набора жестов, представленных в [4], добавив к ним дополняющие жесты из [3]. Но, при демонстрации прототипа словаря представителям общины глухих Москвы, оказалось, что многие жесты, представленные в [4], не совпадают с жестами, используемыми в Москве для обозначения тех же понятий. При этом была высказана просьба не включать в словарь варианты жестов, чтобы «не было путаницы».

В настоящее время словарь формируется на основе наименований жестов, представленных в [4], но выполненных в манере, принятой в Москве. При этом создаваемый словарь не является нормативным, скорее это дескриптивный словарь, описывающий московский диалект русского жестового языка. Но не следует забывать, что литературный русский язык возник на основе московского диалекта…

Учитывая, что большинство предполагаемых будущих пользователей словаря – молодые люди, школьники, дальнейшее пополнение словаря должно вестись на основе современной лексики. В качестве источника этой лексики избран сайт http://youngwriters.ucoz.ru/, на котором представлены литературные произведения школьников и студентов младших курсов институтов, объединенных молодежным пресс-центром «Метаморфозы» московского района Строгино.

Содержание поля «Пояснение к слову»

Пояснения к словам должны передавать смысловые оттенки слов и особенности их употребления для лучшего усвоения норм русского языка глухими, для которых русский язык во многих случаях может считаться вторым языком [2]. Показанные на рис. 3 и 4 примеры, заимствованные, соответственно, из словаря В. Даля и Большой советской энциклопедии, представленных в разделе «Словари» портала Яндекс (www.yandex.ru), свидетельствуют, что тексты пояснений должны тщательно редактироваться, чтобы быть понятным пользователям, чей активный словарь достаточно ограничен. В то же время эти пояснения должны, по возможности, передавать существенную информацию, чтобы способствовать развитию кругозора пользователей словаря. Наиболее эффективно решение этих противоречивых задач может быть осуществлено при участии педагогов школ глухих и слабослышащих и психологов, специализирующихся на работе с глухими.

014-Voskresenskiy_AL:_Layout 1 18.05.2008 14:30 Page 95

Содержание поля «Пояснение к жесту»

В настоящее время поясняющие тексты к жестам взяты из словаря [3]. Но примеры, показанные на рис. 1 и 5, свидетельствуют, что эти пояснения (во многих случаях отсутствующие) слишком кратки и не выполняют возлагаемой на них роли. В связи с этим заполнение этого поля словарных статей, описывающих жесты, ведется с привлечением носителей русского жестового языка. При этом особое внимание уделяется случаям, когда одному слову русского языка соответствуют несколько жестов, передающих различные оттенки смысла. Например, слову «*этаж*» соответствуют различные жесты «*этаж*», соответствующие в одном случае описанию одного из этажей многоэтажного дома, в другом – указывающие, что речь идет о многоэтажном доме.

Использование толковых статей словаря для поиска семантических связей

Семантическая связь лексем и жестов осуществляется путем включения в записи таблиц Words и Signs полей, названных Sem. В настоящее время данные в эти поля заносятся вручную. Выборка нужного жеста (ссылка на соответствующий мультимедийный файл) производится из записи таблицы Signs, имеющей совпадающие значения полей Name и Sem со значениями полей Lexema и Sem таблицы Words, соответственно. В дальнейшем, при накоплении лексикографических данных, заполнение полей Sem планируется автоматизировать, при этом предполагается использовать методы дискриминантного анализа для выделения слов и словосочетаний, наиболее значимых в данной предметной области для определения границ смысловых полей.

Так, например, для приведенных на рис. 5 – 6 примерах, можно считать, что для лексемы «лук (растение)» такими словами будут *«расти», «сорвать», «срезать», «посадить», «острый»*. Эти термины относятся к таксономическому классу «*растение*». Для лексемы «*лук (оружие)*» такими словами будут *«натянуть», «взять», «согнуть», «тугой»*. Эти термины относятся к таксономическому классу «*инструмент*». Конечно, эти примеры являются упрощенными. Для хранения фраз, словосочетаний и отдельных слов, наиболее свойственных контекстам, в которых проявляются семантические значения слов и наименований жестов, в словаре имеются скрытые поля, недоступные пользователю.

Поиск близких по значению записей таблиц Words и Signs предполагается вести с помощью двух процедур. Первая заключается в поиске в полях, содержащих примеры контекстов, одних и тех же слов, однозначно определяющих соответствующий таксон (это может быть и одно слово, встречающееся в соответствующем контексте единично, но с вероятностью, весьма близкой к единице). Если эта процедура не приводит к успеху, то определяются записи таблиц Words и Signs, контекстные поля которых содержат, соответственно, наибольшее число повторяющихся слов и словарных групп.

При использовании одних и тех же тезаурусов для анализа поясняющих текстов и типовых контекстов лексем и наименований жестов можно использовать слова-вершины деревьев отношений «гипероним – гипоним» в качестве наименований таксономических классов.

Дополнение словаря средствами поддержания отношений «часть – целое», «гипоним – гипероним», «синоним – антоним» как для лексем и наименований жестов, так и для словоформ поясняющих статей (т.е. включение в него тезауруса, а позднее развитие этого тезауруса до уровня онтологии) позволит автоматизировать отнесение лексем и наименований жестов к соответствующим таксономическим классам. Соответственно, будет автоматизировано и заполнение полей Sem. Этот подход, как мы надеемся, позволит автоматически тематически связать огромное множество слов со значительно меньшим множеством жестов, используя для этого существующие тезаурусы русского языка, в том числе и русскую версию WordNet.

Заключение

Качественное лексикографическое описание лексем и жестов является весьма критичным для развития создаваемого словаря. Однако, даже из приведенных примеров (рис. 1 – 6) видно, что имеющиеся в настоящее время в словаре тексты пояснений не могут быть признаны удовлетворительными.

Пояснения к словам или слишком кратки (не включают описаний особенностей применения слова) или (при заимствовании из энциклопедий) слишком подробны, но при этом во многих случаях не понятны для лиц, плохо знакомых с русским языком. Кроме того, как выяснилось в ходе работы, для ряда слов (в основном глаголов) отсутствуют (или не найдены) толкования. В ряде случаев толкования даются путем сопоставления с синонимами.

Пояснения к жестам, приведенные в [3], или очень кратки, или вместо пояснения приводится фраза «жесты отличаются по смыслу», при этом не указывается, в чем заключаются различия, каковы особенности применения того или иного жеста. В [4] толкования жестов, которые можно извлечь из примеров применения

014-Voskresenskiy_AL:_Layout 1 18.05.2008 14:30 Page 96

жестовых фраз, весьма ограничены, для многих жестов отсутствуют. Из-за различия в содержании словарей [3] и [4] во многих случаях поле «Пояснение к жесту» в настоящее время не заполнено (см. рис. 6).

Дальнейшим развитием словаря является включение в него тезаурусных отношений для слов и наименований жестов с целью автоматизации разделения смысловых оттенков. Этот шаг необходим для перехода в дальнейшем к обработке не только отдельных слов, но также и фраз или фрагментов текстов, т.е. для перехода к системе автоматизированного перевода текста в жесты.

Список литературы

1. Базоев В.З., Паленный В.А. Человек из мира тишины // М.: Академкнига, 2002.

2. Зайцева Г.Л. Дактилология. Жестовая речь: Учебное пособие для ВУЗов // М.: Просвещение, 1991.

3. Фрадкина Р.Н. Говорящие руки: Тематический словарь жестового языка глухих России // М., 2001.

4. Специфические средства общения глухих // СПб – Павловск: МЦР, 2002. Видеокурс: В 3 частях.

5. Воскресенский А.Л., Хахалин Г.К. Средства семантического поиска // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.). М.: Изд-во РГГУ, 2006. С. 100 – 104.

6. Акуленко В.В. О «Ложных друзьях переводчика» // Электронный документ: http://linguistic.ru/index.php?id=79&op=content

7. Мальковский М.Г. Диалог с системой искусственного интеллекта // М.: МГУ, 1985.

8. Зализняк А.А. Грамматический словарь русского языка // М.: Русский язык, 1980.

9. Селегей В.П. Электронные словари и компьютерная лексикография // AINEWS. Новости искусственного интеллекта. 2001. № 1 (49). Электронный документ: http://www.lingvoda.ru/transforum/articles/pdf/selegey_a1.pdf.

015-Gelbuh_A:_Layout 1 18.05.2008 14:56 Page 97

ГЕНЕТИЧЕСКИЙ АЛГОРИТМ ДЛЯ АВТОМАТИЧЕСКОГО РАЗБИЕНИЯ СЛОВ НА МОРФЕМЫ

GENETIC ALGORITHM FOR AUTOMATIC DIVISION OF WORS INTO MORPHEMES*

Гельбух А.Ф.[1] *(www.gelbukh.com),* ***Сидоров Г.О.***[1] *(www.cic.ipn.m/~sidorov),* ***Лара-Рейес Д.***[1]*,*
Чанона-Эрнандес Л.[2]*,* ***Чубукова М.В.***[3] *(licht66@mail.ru)*
[1]*Лаборатория естественного языка и обработки текста,*
Центр Компьютерных Исследований (CIC),
Национальный Политехнический Институт (IPN), г. Мехико, Мексика
[2]*Инженерный факультет (механика, электричество) (ESIME),*
Национальный Политехнический Институт (IPN), г. Мехико, Мексика
[3]*Кафедра филологического образования,*
Московский институт открытого образования,
Москва, Россия

В статье обсуждается независимая от примеров (unsupervised) техника определения морфемной структуры слов во флективных языках на примере испанского языка. Мы используем глобальную оптимизацию, реализованную с применением генетического алгоритма, без каких-либо эвристик или допущений, уменьшающих размерность задачи a priori. Приводится описание алгоритма. Дается предварительная оценка результатов. Данные на входе представляют собой список слов, построенный на основе корпуса или словаря, а данные на выходе – список тех же слов, разделенных на морфемы. Как и многие автоматические методы, такой алгоритм не претендует на нахождение стопроцентно точного решения и требует ручной постобработки. Тем не менее он позволяет быстро обнаружить тенденции в данных и получить предварительные результаты без больших затрат ручного труда.

1. Введение

В статье обсуждается независимая от примеров (*unsupervised*) техника определения морфемной структуры слов во флективных языках на примере испанского языка. Мы используем глобальную оптимизацию, реализованную с применением генетического алгоритма. Отличие от предыдущих подходов состоит в том, что предыдущие подходы основаны на некотором наборе дополнительных эвристик и предположений, посчитанных для условных вероятностях частей слов, для уменьшения размерности задачи (Goldsmith, 2001). Мы же используем более простую модель, в которой не требуются дополнительные эвристики, применение которых отбрасывает многие возможные варианты решения, которые могут быть правильными. Кроме того, с эвристиками существует риск того, что метод окажется «запертым» в локальном максимуме, поскольку другие возможности просто не рассматриваются. Естественно, у такого более универсального подхода есть и свои недостатки, скажем, его труднее сделать применимым к конкретной задаче, или он работает больше времени.

Напомним, что генетический алгоритм это метод оптимизации (часто применяемый, например, в искусственном интеллекте), взявший за основу модель эволюции, когда лушее решение «выводится», комбинируя свойства индивидов и отбирая лучших индивидов для последующей обработки.

На самом деле, применение методов, независимых от примеров (*unsupervised*), оправдано для определенных классов задач, когда нет возможности решить задачу полным перебором, а применение эвристик требует дополнительных допущений. Тогда можно попытаться сформулировать задачу в форме, подходящей для какого-либо метода, независимого от примеров (*unsupervised*), это может быть генетический алгоритм, или моделирование кристаллизации (*simulated annealing*) или оптимизация потока частиц (*particle swarm optimization*), и пр. То есть, речь не о том, что это «модные» методы, или что все задачи должны решаться с их помощью, а о том, что

* Work done under partial support of Mexican Government (CONACyT, SNI) and National Polytechnic Institute, Mexico (SIP, COFAA, PIFI).

015-Gelbuh_A:_Layout 1 18.05.2008 14:56 Page 98

для некоторые задачи, которые неразрешимы обычными методами, могут быть разрешены с использованием таких методов оптимизации.

В наших экспериментах мы используем испанский язык, который имеет довольно простую морфологическую структуру, поэтому в алгоритме учитывается не более трех возможных морфем. Окончания мы не рассматриваем, отделяя их на этапе предобработки. С некоторыми изменениями наш алгоритм применим к любому флективному языку, в том числе и к русскому.

Заметим также, что мы работаем в рамках словообразовательного подхода, который отличается от более традиционного словоизменительного подхода, заостряющего внимание только на основах и флексиях.

Основные идеи подхода, основаного на эвристиках, описаны в работе (Goldsmith, 2001). Различные вариации этого метода приведены, например, в (Baroni et al., 2002), (Rehman and Hussain, 2005), (Creutz, 2003), см. также (Creutz and Lagus, 2007). Кроме того, было проведено два конкурса по автоматическому разделению слов на морфемы. В 2005 (Pascal challenge, 2005), вышеупомянутые методы были применены к разбиению слов на морфемы на различных языках (английский, финский язык, турецкий язык); испанский язык не рассматривался. В 2007 году (Pascal challenge, 2007), формулировка проблемы была изменена с целью обнаружения разных аспектов значения морфем, что, естественно, вызвало гораздо большие трудности.

Интересные идеи, связанные с улучшением автоматически построенных моделей, приводятся в (Schone and Jurafsky, 2000). Они основаны на использовании повторяющихся контекстов в корпусе. В нашей работе, поскольку мы работаем со словарем, эти идеи не применимы, однако их полезно учитывать в случае работы с корпусом текстов.

Одна из важных идей при автоматическом разбиении слов на морфемы связана с повторяемостью групп морфем. Часто результаты такой группировки называют сигнатурой (подписью), мы же предпочитаем сразу использовать более лингвистический термин - парадигма; в нашем случае это будет деривационная парадигма. Как это и принято, данный термин указывает на тот факт, что корни могут быть связаны с наборами аффиксальных морфем, и эти наборы повторяются, например, в английском, *high, highly, highness* и *bright, brightly, brightness* имеют общий набор суффиксов {∅, *-ly*, *-ness*}. Необходимо заметить, что использование таких наборов очень существенно сокращает пространство поиска алгоритма. Во всяком случае, это верно для испанского и английского языков. Словообразование русского языка гораздо более прихотливо (Кузнецова, Ефремова, 1986), (Тихонов, 2003) и, вероятно, для русского надо учитывать пересечение таких наборов. Скажем, было не так просто найти примеры полного совпадения парадигм, например, *белый* и *серый* с деривационной парадигмой *{-е(ть), -и(ть), -еньк, -оват, -от, -ёхоньк, -ость}*.

Рассмотрим более подробно подход, основаный на эвристиках (Goldsmith, 2001). Метод состоит из двух фаз: разбиение слов на морфемы без парадигм (signatures) и определение парадигм. В первой фазе используются условные вероятности морфем на основе взятого несколько «с потолка» распределения вероятностей (распределение Больцмана, которое давало результаты, более соответствовавшие интуиции автора) для введения метрики в пространстве поиска. Эта метрика позволяет применить итеративный, быстро сходящийся алгоритм оптимизации. Заметим, что в первой фазе этого метода не используется идея MDL (minimum description length), которая используется в его второй фазе для построения лучшего набора парадигм.

В нашем методе нас интересовало «как можно более *unsupervised*» обучение – мы хотели по возможности избежать подобной подгонки параметров под ответ. Плата за это была двоякой. Во-первых, чем более «*unsupervised*» будет алгоритм, тем менее красивые результаты он даст – однако тем более интересен тот факт, что он все-таки дает хоть какие-то результаты. Во-вторых, по сравнению с (Goldsmith, 2001) мы потеряли столь удобную меру на пространстве поиска, и теперь вынуждены для поиска экстремума честно перебирать все варианты (генетическим алгоритмом или любым другим подобным методом «почти полного перебора»). На этой же фазе мы оптимизируем один из вариантов MDL в качестве функции оценки, которая в указанной работе используется вовсе не для этой цели.

С другой стороны, генетические алгоритмы уже применялись для подобных задач (Kazakov, 1997), (Gelbukh *et al*., 2004). Заметим, что в данной статье мы видоизменяем идею, предложенную в этих работах, учитывая повторения парадигм. Это позволяет сократить пространство поиска, но не за счет введения другой метрики, то есть сама задача не видоизменяется.

Как и многие автоматические методы, рассмотренный алгоритм не претендует на нахождения стопроцентного решения и требует ручной постобработки, но он позволяет быстро обнаружить тенденции в данных и получить предварительные результаты без больших затрат ручного труда.

Далее в статье сначала мы описываем алгоритм и его параметры и представляем предварительные экспериментальные результаты.

2. Алгоритм

В этом разделе статьи мы представляем описание алгоритма и обсуждаем его параметры, используемые в экспериментах. Как во всяком генетическом алгоритме, самым важным является представление данных и реше-

015-Gelbuh_A:_Layout 1 18.05.2008 14:56 Page 99

ние о том, что должна вычислять функция оценки (fitness), определяющая, насколько хорош каждый индивид (от этого будет зависеть, останется ли он в следующих популяциях или нет).
Алгоритм содержит следующие шаги:

1. Для каждого слова, мы обнаруживаем все потенциально возможные приставки, начиная от первой буквы и заканчивая предпоследней, включая пустую приставку ∅. В список возможных морфем добавляются возможные приставки без повторений с их частотами.

2. Готовится список возможных корней, начинающихся от первой буквы и кончая последней. Список корней содержит уникальные элементы с их соответствующими частотами и не включает пустой корень ∅.

3. Строится список возможных суффиксов, который представляет собой наборы, организуемые от последней буквы слова до второй буквы от его начала. Этот список содержит пустой суффикс ∅ и также не разрешает повторений. Частоты суффиксов фиксируются.

4. Полученный список фильтруется следующим образом. Для корней обнаруживаются все возможные парадигмы, то есть, наборы морфем, которые повторены несколько раз для различных корней. Например, парадигмы {NULL, *-ism, -iz, -idad*}, {NULL, *-ant, -acion*}. Мы отфильтровываем все некорневые подмножества (потенциальные морфемы), которые содержат только один элемент. Кроме того, мы отфильтровываем морфемы, которые принадлежат только парадигмам с низкой частотой. В нашем случае, мы использовали в качестве порога частоту равную трем, то есть морфема должна входить по крайней мере в одну парадигму с частотой четыре и больше (то есть повторится по крайней мере у четырех слов).

5. Формируются хромосомы в соответствии со следующим правилам:
   a. Хромосомы являются бинарными (двоичными), в том смысле, что они состоят из генов, которые двоичны. А хромосома это набор генов.
   b. Каждый ген (двоичное место в хромосоме) соответствует элементу одного из списков,
   c. Длина хромосомы это сумма числа элементов трех списков.
   d. Значение гена «1» в хромосоме означает, что соответствующий элемент соответствующего списка является частью решения, в то время как значение «0» указывает на то, что этот элемент списка не участвует в словообразовании.

6. Порождается начальная популяция, состоящая из нескольких индивидов (хромосом), в генах которых случайным образом расставляются нули и единицы.

7. Применяется генетический алгоритм с различными значениями параметров (см. ниже). Для каждой вновь полученной хромосомы вычисляется функция оценки для решения о том, останется ли она в следующей популяции или нет. Заметим, что если мутация или скрещивания производят хромосому, которая «неправильна», то есть содержит морфемы, которые не существуют в отфильтрованных списках, тогда хромосома «улучшается» путем добавления случайным способом аффиксов, исправляющих эту «дефектную» хромосому. В худшем случае, такое слово добавляется в список корней с нулевыми аффиксами. Заметим, мутация такова, что нулевые аффиксы никогда не выключаются.

8. Функция оценки каждой хромосомы вычисляется следующим образом:

a. Общее число генов, используемых в решении (гены отличные от нуля), должно быть как можно меньше. Заметим, что мы уже обеспечили покрытие всего списка слов, когда «улучшали» хромосому в случае необходимости.

b. Частоты используемых элементов должны быть как можно больше, то есть высокочастотные элементы премируются. Под частотой здесь имеются в виду участие элемента в разбиении разных слов, то есть чем в большем числе слов он встречается, тем лучше.

В данной версии алгоритма мы делаем это по формуле:

$$fitness = -\log(freq(x0)) - \log(freq(prefix) * freq(stem) * freq(suffix))$$

где

freq(x0)) - сумма всех позиций в хромосоме, где стоят нули,

freq (prefix) - суммарная частота всех префиксов, участвующих в решении (их ген равен 1),

freq (stem) - суммарная частота всех корней, участвующих в решении (их ген равен 1),

freq (suffix) - суммарная частота всех суффиксов, участвующих в решении (их ген равен 1).

Поскольку минимизируется общий размер списка элементов, то можно рассматривать данную функцию как одну из разновидностей идеи minimum length description (MDL).

9. По окончании работы алгоритма из последней популяции выбирается хромосома с наибольшим значением функции оценки. Она и является наилушим, с точки зрения алгоритма, разбиением слов исходного словаря на морфемы.

015-Gelbuh_A:_Layout 1 18.05.2008 14:56 Page 100

Необходимо упомянуть следующие традиционные параметры генетического алгоритма:

1. Замещение и выбор родителей. В нашем случае, генетический оператор выбора родителей производится путем использования схемы турнира, а именно, для двух случайно выбранных индивидов сравнивается их пригодность (на основе функции оценки). Это гарантирует, что индивиды конкурируют и что выживают лучшие. Замещение определяет процент индивидов (хромосом), которые должны быть заменены в популяции.

2. Скрещивание. Генетический оператор скрещивания использует в качестве параметра число блоков, на основе которых хромосомы обмениваются своей генетической информацией для создания новых хромосом. Места для скрещивания выбираются случайным образом.

3. Наследование. Оперотор наследования определяет, какие индивиды остаются в популяции. Мы использовали схему наследования с элитизмом, при которой лучшие индивиды всегда сохраняются в популяции.

4. Мутация. Этот генетический оператор изменяет значения случайно выбранных генов с определенной вероятностью. Этот оператор важен, потому что он позволяет рассматривать новые возможности в пространстве решений.

5. Количество поколений. Этот параметр определяет, сколько раз должен применяться генетический алгоритм для популяции (для каждого индивида (хромосомы) в популяции).

6. Размер популяции. Этот параметр определяет, сколько индивидов одновременно существует в популяции. Обычно их должно быть не меньше 100.

Мы провели эксперименты с разными параметрами генетического алгоритма и сравнили их результаты. Выяснилось, что лучшие результаты получаются при использовании трех различных наборов параметров, то есть, алгоритм применяется последовательно три раза с разными параметрами.

Мы экспериментировали в пределах следующего диапазона параметров, см. Таблицу1.

	Минимальные параметры	Максимальные параметры
Размер популяции	50	5,000
Замещение	20%	100%
Мутация	Начинается с 20 %, уменьшается в соответствии с числом поколений	Начинается с 90 %, уменьшается в соответствии с числом поколений
Скрещивание (количество блоков, где происходит скрещивание)	1	20
Поколения	50	10,000

Таблица 1. Диапазоны параметров генетического алгоритма

Лучшие результаты были получены для следующих последовательно примененных наборов параметров генетического алгоритма для популяции в 200 индивидов; см. Таблицы 2, 3 и 4.

Замещение	60%
Мутация	30%, уменьшается в соответствии с числом поколений
Скрещивание	5
Поколения	10,000

Таблица 2. Параметры первого прохода

Замещение	80%
Мутация	80%, уменьшается в соответствии с числом поколений
Скрещивание	20
Поколения	7,000

Таблица 3. Параметры второго прохода

015-Gelbuh_A:_Layout 1 18.05.2008 14:56 Page 101

Замещение	80%
Мутация	80%, уменьшается в соответствии с числом поколений
Скрещивание	20
Поколения	7,000

Таблица 4. Параметры третьего прохода

Замещение 40% Мутация 20%, уменьшается в соответствии с числом поколений Скрещивание 4 Поколения 6,000

Цели каждого из проходов различны. При первом проходе популяция подготавливается и упорядочивается. При втором проходе происходит максимальная «встряска» популяции. Наконец, при третьем проходе популяция должна стабилизироваться, например, значительно уменьшена норма мутации.

3. Экспериментальные результаты

В качестве входных данных мы использовали испанский словарь, который содержал более 20,000 значимых слов. Мы игнорировали вспомогательные слова и наречия, работая только с существительными, глаголами и прилагательными. Так как мы интересуемся словообразовательной морфологией, то мы не рассматриваем окончания и заранее их отбрасываем, например, *, trabajar* → *trabaj- (работа(ть))*, *rojo* → *roj- (красн(ый))*, и т.д. Кроме того, мы не различали акцентуированные и неакцентуированные гласные, потому что акцент в испанском языке имеет чисто орфографическую функцию. Заключительный входной список содержал 16,849 уникальных слов без окончаний.

Для того, чтобы иметь возможность сравнивать наши данные с одной из самых известных систем - системой Linguistica (Goldsmith, 2001) - в настоящий момент мы провели эксперименты только для отделения суффиксов, хотя алгоритм позволяет производить одновременную обработку суффиксов и приставок.

Обработка наших данных системой (Gelbukh et al., 2004), которая достаточно успешно применяется к словоизменительной морфологии на корпусе, не дала положительных результатов по причине низкой частотности словообразовательных аффиксов. Полученная точность была менее 10%. Это показывает важность применения парадигм для данной задачи.

При подготовке списка начальных наборов для алгоритма были получены следующие результаты. Были выделены 7,747 деривационных парадигм, из которых 6,472 содержали более чем один элемент. Парадигмы, которые содержали ровно один элемент, были отфильтрованы. Стоит упомянуть, что из этих 6,472 парадигм 1,852 парадигмы содержали нулевой суффикс.

Дополнительно, мы использовали порог на повторяемость парадигмы, а именно, мы игнорировали парадигмы, которые повторились меньше чем три раза, то есть, они существуют для трех слов или меньше. Всего было 5,535 парадигм с частотой равной «1», 404 с частотой, равной «2» и 171 с частотой равной «3». В конце концов, только 372 парадигмы обрабатывались алгоритмом, то есть только суффиксы и корни, которые участвовали в них, были использованы для представления хромосом. В результате алгоритм работал с 17,085 корнями и 136 суффиксами. Это очень существенное уменьшение пространства поиска, если учесть, что вначале число потенциальных корней было более 44,000, а потенциальных суффиксов – более чем 15,000.

К сожалению, золотого стандарта разбиения на морфемы для испанского языка не существует (или мы его не нашли), и мы планируем его составить. К моменту написания этой статьи мы смогли провести оценку только на небольшой выборке из полученных результатов (можно считать его крохотным золотым стандартом, который в будущем будет расширен). А именно, мы сравнили наши результаты с результатами, полученными с применением системы Linguistica, для одних и тех же входных данных. Из нашего словаря было взято 5,000 первых слов. Это число связано с тем, что доступная версия системы Linguistica принимает на входе не больше этого количества слов. Обе системы обработали эти данные и получили результаты разбиения слов. В этих результатах мы взяли первые 100 слов из каждого из них и проверили вручную правильность разбиения. Система Linguistica получила 87% точности, а наша система 84%. Заметим, что это предварительные данные которые показывают, что наша система производит сопоставимое с существующими разбиение слов на морфемы, полная оценка может измениться в ту или другую сторону. Мы оценивали только точность, а не точность, полноту и F-меру, поскольку в наших экспериментах мы определяли только суффиксы, то есть точка разбиения была только одна на слово (а не несколько, как, напр., у (Creutz and Lagus, 2007), поэтому все три меры (P, R, и F) совпадают. В будущем мы будем определять как приставки, так и суффиксы, и тогда потребуется применять эти три меры. Конкретные разбиения слов могут не совпадать, например, система Linguistica не находит суффикс *–miento),* который был довольно частотен в списке и был обнаружен нашей системой. Точная оценка полученных результатов дело будущих исследований.

015-Gelbuh_A:_Layout 1 18.05.2008 14:56 Page 102

4. Выводы

Мы описали применение достаточно универсального оптимизационного метода, а именно, генетического алгоритма, к проблеме разбиения слов на морфемы на уровне словообразования. Наши эксперименты были проделаны на материале испанского языка.

Полученные результаты сопоставимы с результатами методов, основанных на вычислениях вероятностей с использованием эвристик для уменьшения размерности задачи. Нам представляется важным, что для решения нашей задачи мы не пользуемся дополнительными эвристиками, которые могут давать неплохие результаты, но при этом могут изменить саму формулировку проблемы, как, скажем, предположение о распределении Больцмана для условных вероятностей морфем в наиболее широко используемом подходе.

С другой стороны, очевидно, что решить проблему разбиения слов на морфемы методом полного перебора невозможно. Нам представляется важным представить возможность решения такой проблемы методом «почти полного перебора», с использованием одного из методов обучения без примеров (unsupervised).

Хромосома для алгоритма строится из ряда возможных наборов для корней и аффиксов, отфильтрованных специальным образом. Хромосомы являются двоичными, где «1» отмечает присутствие элементов в возможном решении и «0» - их отсутствие. Алгоритм учитывает деривационные парадигмы, т.е. повторяющиеся наборы аффиксов. Мы использовали в качестве фильтра частотность парадигмы (по крайней мере, четыре слова должны иметь такую парадигму). Также мы игнорировали парадигмы, которые состояли только из одного элемента.

Традиционные операторы генетического алгоритма были применены к обработке хромосом (=индивидов) в популяциях.

В настоящий момент, мы отфильтровываем все некорневые наборы (потенциальные морфемы), которые не являются частью какой-либо дервивационной парадигмы. В будущем, мы собираемся производить обработки этих парадигм иным способом, а именно, включить их в функцию оценки или включить их непосредственно в хромосомы.

Будущие исследования также связаны с выполнением точной оценки результатов разбиения слов на морфемы для испанского языка. Заметим, что для испанского языка не существует золотой стандарт в это области. Мы планируем разработать этот стандарт.

Было бы интересно просести аналогичные исследования для русского языка, где такой стандарт есть (Кузнецова, Ефремова, 1986). В случае русского языка, как мы уже упоминали, видимо, необходимо по-другому трактовать парадигмы, рассматривая их пересечения.

Список литературы

1. Baroni M, Matiasek J, Trost H. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. // ACL Workshop on Morphological and Phonological Learning. 2002.

2. Creutz M. Unsupervised Segmentation of Words Using Prior Distributions of Morph Length and Frequency. // Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03), Sapporo, Japan, July 2003, pp 280-287.

3. Creutz M., Lagus K. Unsupervised models for morpheme segmentation and morphology learning. // ACM Transactions on Speech and Language Processing (TSLP), January 2007, v.4 n.1, 34 p.

4. Gaussier E. Unsupervised learning of derivational morphology from inflectional lexicons. // Proceedings of the ACLWorkshop on Unsupervised Learning in Natural Language Processing. University of Maryland. 1999. pp 24–30.

5. Gelbukh A., Alexandrov M, SangYong Han. Detecting Inflection Patterns in Natural Language by Minimization of Morphological Model. // A. Sanfeliu, J. F. Martínez Trinidad, J. A. Carrasco Ochoa (Eds.). Lecture Notes in Computer Science N 3287, Springer-Verlag, 2004, pp. 432–438.

6. Goldsmith J. Unsupervised Learning of the Morphology of a Natural Language. // Computational Linguistics 27:2 (2001) pp. 153-198.

7. Haahr P., Baker S. Making search better in Catalonia, Estonia, and everywhere else // Google, 2007 http://googleblog.blogspot.com/2008/03/making-search-better-in-catalonia.htm

8. Kazakov D. Unsupervised learning of naıve morphology with genetic algorithms. // Workshop Notes of the ECML/MLnet Workshop on Empirical Learning of Natural Language Processing Tasks. Prague, Czech Republic, 1997, pp. 105–112.

9. Pascal Morphochallenge // 2005. http://www.cis.hut.fi/morphochallenge2005/

10. Pascal Morphochallenge // 2007. http://www.cis.hut.fi/morphochallenge2007/

11. Rehman Kh, Hussain I. Unsupervised Morphemes Segmentation. // Pascal Morphochallenge, 2005, 5 p.

12. Schone P., Jurafsky D. Knowledge-free induction of morphology using latent semantic analysis. // Proceedings of the Conference on Computational Natural Language Learning (CoNLL-2000), 2000.

13. Snover M.G., Brent M.R. A Bayesian model for morpheme and paradigm identification // Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Toulouse, France: ACL, 2001, pp. 490 – 498.

14. Snover M.G., Brent M.R.A Probabilistic Model for Learning Concatenative Morphology. // Proceeding of NIPS (Neural Information Processing Systems), 2002.

15. Кузнецова А.И, Ефремова Т.М. Словарь морфем русского языка. // М.: Рус.яз., 1986 -с.1136

16. Тихонов А.Н. Словообразовательный словарь русского языка: В 2 т. // М.: ООО «Издательство Астрель»; «Издательство АСТ», 2003. т1.- 860 ; т.2 -941.

016-Gerassimenko_2:_Layout 1 18.05.2008 15:01 Page 103

ДВА ЗНАЧЕНИЯ – ДВЕ ЯЗЫКОВЫХ ЕДИНИЦЫ? *АГА* В СПОНТАННОМ ДИАЛОГЕ*

TWO MEANINGS, TWO LINGUISTIC ITEMS? RUSSIAN „*AHA*" IN SPONTANEOUS DIALOGUE

Герасименко О.А. *(olga.gerassimenko@ut.ee)*
Тартуский университет, Эстония

Рассматриваются примеры использования прагматической частицы *ага* в спонтанных устных институциональных диалогах и обсуждается возможность нахождения общего семантического знаменателя у языковых единиц, выражающих, в зависимости от секвенциальной позиции, согласие/подтверждение или удивление/удовлетворение.

1. Введение

Лексема *ага* относится к прагматическим частицам, которые помогают участникам диалога осуществлять обратную связь и, тем самым, лаконично и эффективно управлять ходом диалога. Закономерности использования тех или иных прагматических частиц зачастую не осознаются говорящим, но могут разниться для представителей разных культур, поэтому описание функционирования каждой частицы важно для изучающих язык. В современных исследованиях русских междометий лексема *ага*, как правило, рассматривается в совокупности с другими частицами: в группе междометий слушающего, выполняющих контроль за процессом восприятия информации (наряду с *угу* и *гм*, Шаронов 2005: 212), выражающих пассивную реакцию слушающего (наряду с *угу*, Добрушина 1998: 138-139) или реакцию на окончательное осознание чего-либо (наряду с *а* и *ах*, Шаронов 2006); в числе междометий догадки, сопровождающих мыслительный процесс и произносимых в его критические моменты (наряду с *а, о, э, эге*, Иомдин 2004).

В словарных же описаниях, наряду с междометием *ага*, выражающим «догадку, радостное удивление, злорадное торжество», мы встречаем утвердительную частицу *ага*, выражающую «согласие, подтверждение» (ССРЛЯ 1991). Этого же разграничения придерживается Анна Вежбицкая, рассматривая *ага* в числе когнитивных междометий: она отмечает, что два значения *ага*, согласие/подтверждение и удовлетворение/удивление, не могут быть объединены инвариантом, и предполагает, что эти значения выражаются омонимичными языковыми единицами (Wierzbicka 1991: 329). Интересно, что словарь Ожегова и Шведовой (ТСРЯ 1997), сохраняя различение междометия и частицы, называет междометием «восклицание с торжествующей интонацией», а в значения частицы, помимо подтверждения, заносит также «уяснение или догадку», относимые другими источниками к междометию. В текстах со снятой грамматической омонимией Русского национального корпуса употребления лексемы размечены как междометия (INTJ), если речь идет о внезапной догадке (прим. 1) или принятии к сведению (2), и как частицы (PART), если *ага* выражает подтверждение (3) или согласие (4).

(1) INTJ
Па́пка, с кото́рой пришёл Ганка, лежа́ла на столе́. Краси́вый офице́р взял её в ру́ки, полиста́л немно́го и сказа́л: «*Ага́*!» Он сказа́л «*ага́*» таки́м то́ном, кото́рый зна́чил: «Так вот заче́м вы сюда́ собра́лись».

(2) INTJ
- Ста́рый слуга́ семе́йства Курцеров, - сказа́л Ланэ. -С господи́ном полко́вником он встреча́лся ещё в Чехии. Они́ там вме́сте в како́й-то лаборато́рии рабо́тали. -*Ага́*, так? - при́нял к све́де́нию Гарднер.

(3) PART
-Нет, Андреич, не до ча́ю, е́ду на свида́ние. -Здорово! - обра́довался Потапов. -Со стару́шкой? -*Ага́*.

(4) PART
- А чего́-то оно в го́рле остана́вливается... Ни у кого́ не остана́вливается? - Да, распира́ет ка́к-то. - *Ага́*!.. И в нос бьёт!

* Работу поддерживает Эстонский научный фонд (гранты № 7503 и 5813).

016-Gerassimenko_2:_Layout 1 18.05.2008 15:01 Page 104

Однако есть и случаи, в которых нейтральное принятие к сведению размечено как частица (5), а ироническое псевдосогласие как междометие (6), а также контексты, в которых снятие омонимии представляется затруднительным, так как лексема может выражать и согласие, и принятие к сведению (7).

(5) PART
Вы Мишу Панина зна́ете? - Нет. -Наш заве́дующий литерату́рной ча́стью. - *Ага́*.

(6) INTJ
- А сего́дня холо́дно? -Жу́ткая холодрыга. - А Маша говори́т тепло́. - *Ага́*/ с но́су потекло́. Э́то твоя́ Маша молода́я / у неё кровь горя́чая.

(7) PART
- Дава́й я тебя́ провожу́, - сказа́л Колюня. Она́ скользну́ла по его лицу́ лени́вым взгля́дом, сощу́рилась и нетерпеливо мотну́ла голово́й, но он вы́держал - лежа́вший на песке́ карп прида́л ему́ сил. - Артур, мы ухо́дим! - сказа́ла Аня. -*Ага́, ага́,* - закива́л он, хло́пая на го́лой спине́ комаро́в. Колюня засу́нул ка́рпа в холщо́вый мешо́к, и ма́льчик с де́вочкой пошли́ домой.

Мы попробуем рассмотреть случаи употребления *ага* в спонтанном устном диалоге и сопоставить диалогические функции *ага* с приведенными выше семантическими описаниями. Обладают ли употребления *ага*, разделяемые на междометное и частичное, некими инвариантными чертами, или же мы имеем дело с омонимами?

2. Корпус и метод

Исследование проведено на материале имеющегося в распоряжении автора небольшого корпуса телефонных регистратурных диалогов (45 минут звучания, 81 диалог, ок. 9000 слов расшифровки), собранных в наиболее изолированном от языкового контакта русскоязычном регионе Эстонии. Регистратурные диалоги отличаются четкой тематической структурой и свойственной институциональному диалогу нейтральностью (однако, и эмоциональность не исключена полностью), что позволяет с большей легкостью сопоставлять контексты использования языковых средств. Запись производилась на телефоне регистратуры и включала все входящие и исходящие звонки (т.е. разговоры пациентов или их близких с регистраторами и рабочие разговоры медиков между собой); о записи знала лишь одна сторона диалога. Все говорящие анонимны, приводимые социальные параметры извлечены из аудиозаписи. Диалоги расшифрованы в транскрипции конверсационного анализа (Jefferson 2004), фиксирующей важные параметры общения: временную соотнесенность реплик, длительность пауз, ударные слова и интонационные единицы (см. приложение).

Конверсационный анализ, исследующий аутентичные ситуации речевого взаимодействия, рассматривает прагматические единицы в процессе ведения диалога, исходя из взаимной интерпретации происходящего участниками диалога, которая проявляется в их последовательных реакциях и временной соотнесенности этих реакций (Hutchby, Wooffitt 1998). Транскрипция используется как вспомогательное средство для поиска интересующих исследователя явлений и представления примеров (Have 1999: 77), анализируется непосредственно запись диалога (в случае телефонного диалога – аудиозапись, полностью воспроизводящая речевую ситуацию).

3. АГА в спонтанной речи

В корпусе содержится 69 употреблений *ага*. В связи с многочисленностью вопросо-ответных секвенций в регистратурном диалоге (звонящий задает вопросы для получения информации о приеме врача или способах связаться с врачом, работник регистратуры при помощи вопросов заполняет слоты регистрации визита) большинство употреблений лексемы приходится на третий член вопросо-ответной секвенции, сигнализирующий собеседнику, что ответ услышан и принят (35 случаев). Меньшее число встречается в схожем контексте т.н. «ремонтных» вопросов и ответов (Schegloff 1992), связанных с разрешением проблем общения: недослышанности, недопонятости (16 случаев). Такие секвенции отличаются тем, что ответ в них чаще всего служит лишь подтверждением (или опровержением) переспрашивающего вопроса и не содержит новой информации. *ага* также встречается в качестве утвердительного ответа на вопрос (4 случая), подтверждения телефонного контакта (1 случай), опознающей реакции на самоидентификацию собеседника (4 случая), положительной реакции на просьбу (3 случая) и на сигнал о завершении диалога (6 случаев).

Рассмотрим два контекста использования *ага*, наиболее различающиеся и наиболее соответствующие словарным значениям междометия и частицы.

016-Gerassimenko_2:_Layout 1 18.05.2008 15:01 Page 105

3.1. Реакция на ответ: принятие к сведению

В качестве реакции на ответ в вопросо-ответных секвенциях *ага* маркирует релевантность полученной информации («уяснение») и часто реагирует непосредственно на релевантную информацию в наложении на продолжающуюся реплику собеседника (8) или «приклеивается» к реплике без естественной паузы между ними (9). Полученный ответ при этом не противоречит ожиданиям говорящего (ср. эхоповтор в (8) и реакцию *а* в (10)), но и не является полностью очевидным (ср. реакцию *угу* в (11)).

(8)
З – мать пациента
О – регистратор, женщина
1. О: `ждите доктора. `будет [{она.}]
2. З: [а] (.) вот до сколь`ки она принимает во `сколько примерно [штобы там]
3. О: [ну она сё]ня работает в `вечер,
4. З: [в `вечер,]
5. О: [значит] будет `до `трёх ча`сов [у вас.]
6. З: [**ага**]
7. О: `пож[алуста.]
8. З: [угу, угу,] `всё, спа`сибо.

Первая в приведенном отрывке диалога реплика О заканчивает регистрацию домашнего визита (предшествующая часть диалога опущена), З прерывает ее прогнозируемое окончание уточняющим вопросом о времени прихода врача. *до скольки* предполагает, что врач ведет прием с утра, однако регистратор опровергает это предположение (строка 3), на что З реагирует эхоповтором релевантной части предыдущей реплики. Прагматическая незавершенность реплики О маркируется незаконченной, слегка понижающейся интонацией (эту интонацию воспроизводит и эхоповтор); свою следующую реплику (строка 5) О оформляет как вывод из предыдущей, выделяя релевантное *до трех часов* ударной интонацией. Непосредственно после этого З произносит нейтрально интонированное *ага*, накладывающееся на окончание предыдущей реплики (строка 6). О истолковывает его как сигнал о достаточности ответа и инициирует завершение диалога (реплика *пожалуйста*), З проявляет понимание и согласие с этим в последней реплике.

(9)
З – медсестра, молодая женщина
О – регистратор, женщина
1. ((звонок))
2. О: [поли]клиника слушает, (.) добрый день.=
3. З: [{алё?}]
4. З: =а `здравствуйте девочки Ма`рина Велесова, [.хх а] Надин Ана`тольевны `нету там?
5. О: [* ага *]
6. О: она=э: (.) пошла=э (.) к секрета`рю.=
7. З: > =**ага**, < а пере`дайте пожалуста, што вот мы там новое распи`сание давали, што Мар`тынова-то за Фа`лангина [да?]
8. О: [угу.]

З – медсестра врача, которая хочет сообщить старшей регистратурной сестре об изменениях в расписании. Она начинает разговор с дружеского приветствия и представления себя (строка 4), на которое регистратор О реагирует негромким опознающим *ага* (строка 5), закрывающим секвенцию рутинной идентификации в наложении на продолжающуюся реплику партнера. В ней З справляется о старшей сестре, с которой, очевидно, хочет поговорить (окончание строки 4); отрицательная форма вопроса может быть как проявлением вежливости, так и проявлением понимания, что старшей сестры может не быть в регистратуре. Ответ О максимально кооперативен и, судя по быстроте реакции З, предсказуем. З принимает его к сведению убыстренным *ага* (строка 7) и переходит к просьбе передать отсутствующей сестре необходимую информацию, попутно напоминая О контекст просьбы.

(10)
З – отец пациентки, молодой мужчина
О – регистратор, женщина
1. О: `ждите доктора.

016-Gerassimenko_2:_Layout 1 18.05.2008 15:01 Page 106

2. З: а ког`да будет?
3. О: .хх (.) ой, так, < `Васина работает сёдня > (0.4) > после `часу ждите. <
4. (.)
5. З: к `часу?
6. О: `после ча[су.]
7. З: [**а**,] `после часу. [ха-ра]шо.
8. О: [* угу *]

Начало отрывка схоже с началом предыдущего; первая часть ответной реплики регистратора (строка 2: вдох, паузы, эмоциональная частица, замедленный темп речи) указывает на то, что информации для ответа нет у нее под рукой, однако она быстро находит ответ и высказывает его в убыстренном темпе. После микропаузы на строке 4 З переспрашивает недослышанную информацию. Он задает общий вопрос в утвердительной форме, предпочитаемым ответом на который является (минимальное) подтверждение; однако, ответ противоречит ожиданиям вопроса, он исправляет неверное понимание и выделяет исправление ударением (строка 6). З демонстрирует понимание этого частицей *а* в наложении на прогнозируемый конец продолжающейся реплики и повторяет исправление, после чего дает информации эмоциональную оценку (раздельно произнесенное *ха-рашо*). О проявляет свое восприятие частицы *а* как сигнала незавершенности секвенции, так как дожидается конца повтора на строке 7 и подтверждает его правильность частицей *угу* (строка 8).

(11)
З – пациент, женщина
О – регистратор, женщина

1. З: как принимает севодня (.) Бого`словский.
2. О: .х Бого`словский севодня с пят`нацати до девят`нацати.
3. З: с `трёх до с:еми, [да?]
4. О: [угу?]
5. З: **угу**, спасибо.

Отрывок из короткого регистратурного диалога начинается вопросом З о времени приема врача, после ответа З уточняет время приема в другой системе измерения. На общий вопрос О отвечает минимальным подтверждением (*угу* с восходящей интонацией), что является ожидаемой реакцией на вопрос: З проявляет понимание этого, произнося частицу *угу* и незамедлительно переходя к благодарности, которая служит также сигналом завершения разговора.

3.2. Ответ на вопрос: подтверждение правильности

В качестве ответа на вопрос *ага* выражает подтверждение и согласие, так же как лексемы *да* (прим. 12, строка 5) и *угу* (прим. 11, строка 4). Однако, *ага*, в отличие от них, в этой позиции встречается редко и именно в тех случаях, когда говорящему почему-либо нужно подчеркнуть правильность догадки собеседника (прим. 12, 13).

(12)
З – медсестра
О – медсестра

1. З: [я горю вы `знаете,] (0.5) не `знаю я горю `девочка не долж`на отвечать. [мхе мхе]
2. О: [как `ты вчера,] мхе {вчера те} `тётенька какая-то чу`жая или `девочка, [не ту`да наверно набирают.]
3. З: [мхе мхе мхе .хх] наверно, .х ну да`вай тогда, [всего.]
4. О: [ну чего] {а} ты с регистра`туры [звонишь, да, там не под]`ключено?=
5. З: [да: да да да]
6. З: =**ага** не [не не]
7. О: [от А]`риши она зво`нила, ни`чё не спрашивала?

З звонит больной коллеге О, чтобы посоветоваться по рабочему вопросу, и, завершив обсуждение вопроса (опущено), со смехом рассказывает, как коллеги тщетно пытались связаться с О (трубку брала какая-то девочка, и З убеждала коллег, что девочка по телефону О отвечать не должна). О с удовольствием поддерживает тему и, когда З подает сигнал к завершению (строка 3), реагирует быстро и разочарованно (*ну чего* в

016-Gerassimenko_2:_Layout 1 18.05.2008 15:01 Page 107

наложении), однако тут же находит объяснение резкому завершению разговора и высказывает гипотезу, ожидающую подтверждения (строка 4). З с энтузиазмом подтверждает обе части версии О, как только их распознает – после выделенных ударением слов *регистратуры* и *подключено*. *ага* «склеивается» с последним словом реплики О и сигнализирует правильное понимание ситуации в целом: по регистратурному телефону нельзя говорить свободно, а «удобный» телефон недоступен. После этого подтверждения О делает попытку обсудить остро интересующий ее вопрос, формулируя его так, чтобы позволить З ответить, не привлекая к предмету разговора внимания окружающих.

(13)
З – врач, пожилая женщина
О – регистратор, женщина

1. О: хх `четыре у вас. [(.) Ваба]`дусе двадцать восемь `тридцать один,
2. З: [четыре, пять,] ((жалобно))
3. (0.7)
4. З: так,
5. О: Ваба`дусе двадцать восемь `семьдесят восемь,
6. З: двадцать восемь `ТРИДЦАТЬ один,=
7. О: =**ага**,=
8. З: =[двадцать восемь]
9. О: [и Ваба`дусе] двадцать восемь `СЕМЬДЕСЯТ восемь.
10. З: `семисят восемь.=
11. О: =**ага**, `Карлова сорок семьдесят `восемь?

Регистратор О перечисляет врачу З адреса домашних визитов на этот день (строки 1-5; в реплике на строке 2 врач жалуется на ежедневную многочисленность визитов). Фиксацию первого адреса и готовность слушать дальше З сигнализирует частицей *так* после паузы, очевидно связанной с записью (строки 3-4). Реплика на строке 5 вызывает затруднение, так как номер дома в двух адресах одинаков. В следующей реплике З переспрашивает первый адрес со слегка нисходящей интонацией продолжения, выделяя голосом номер квартиры (строка 6) и после короткой паузы начинает называть второй адрес (строка 8). О подтверждает правильность первого адреса во время короткой паузы между адресами (строка 7), и одновременно со следующей репликой З повторяет второй адрес (строка 9), выделяя номер квартиры. За этой репликой следует еще один повтор З, на который О реагирует аналогичным подтверждением и называет следующий адрес (строка 11).

4. Выводы и перспективы исследования

При рассмотрении двух позиционных групп употреблений *ага* в устном диалоге выявляются общие черты используемых в них языковых единиц: *ага* реагирует на целостную и релевантную информацию, не являющуюся ни тривиальной, ни противоречащей ожиданиям говорящего. *ага* в качестве реакции на ответ сообщает: «теперь мне понятно», *ага* в качестве подтверждающего ответа несет сообщение «теперь тебе понятно». Нам представляется, что присущая междометию эмоциональность, в зависимости от контекста выражающая удивление, злорадство, догадку или иронию, не отменяет общего семантического компонента, но дополняет его.

Более пристального внимания заслуживают и другие позиционно-функциональные группы употреблений *ага* в регистратурном диалоге, так же как и сопоставление прагматических частиц в нарративных секвенциях, где они связаны не только с членением информационной структуры (как в рассмотренных диалогах), но и с поддержанием канала общения. Безусловный интерес представляет степень формальности лексемы в сопоставлении с другими частицами общего семантического поля. Наконец, было бы интересно узнать, отличается ли употребление прагматических частиц русскоязычными жителями Эстонии от речевых привычек русских из других регионов.

Список литературы

1. Добрушина Н.П. 1998: Семантика междометий в реактивных репликах // Вестник Московского университета 2: 136-145.

2. Иомдин Б.Л. 2004. Междометия догадки в русском языке // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям Диалог’2004. Москва: Наука, http://www.dialog-21.ru/Archive/2004/Iomdin%20B.pdf (использовано 25.01.2008).

016-Gerassimenko_2:_Layout 1 18.05.2008 15:01 Page 108

3. ССРЛЯ = Словарь современного русского литературного языка 1991. Москва: Российская Академия наук, Институт русского языка им. В. В. Виноградова.

4. ТСРЯ = Толковый словарь русского языка 1997. С.И. Ожегов и Н.Ю. Шведова Москва: Российская Академия наук, Институт русского языка им. В.В. Виноградова.

5. Шаронов И.А. 2005. Междометия в речевой коммуникации // Эмоции в языке и речи. Сборник научных статей. Ред. И.А. Шаронов. Москва: РГГУ, 200-220.

6. Шаронов И.А. 2006. О новом подходе к классификации эмоциональных междометий // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям Диалог'2006. Москва: Наука, http://www.dialog-21.ru/dialog2006/materials/html/Sharonov.htm (использовано 25.01.2008).

7. Have, P. ten 1999. Doing Conversation Analysis. A Practical Guide. London: Sage.

8. Hutchby I., Wooffitt R. 1998. Conversation Analysis. Principles, Practices and Applications. Cambridge: Polity Press.

8. Jefferson G. 2004. Glossary of transcript symbols with an introduction // Conversation Analysis. Studies from the first generation. Amsterdam/Philadelpia: Benjamins, 13-59.

9. Schegloff E.A. 1992. Repair after next turn: The last structurally provided defense of intersubjectivity in conversation // American Journal of Sociology, 98: 1295-1345

10. Wierzbicka A. 1991. Cross-cultural pragmatics. The semantics of human interaction. Berlin: Mouton de Gruyter.

Приложение. Знаки транскрипции

звонящий	З:
отвечающий	О:
спад интонации	текст.
полуспад интонации	текст,
подъем интонации	текст?
микропауза (до 0.2 сек.)	(.)
длина паузы в десятых долях секунды	(0.8)
наложение реплики на реплику	[текст]
слияние независимых единиц	=
растянутый звук	а:
ударное слово	`а
прерванное слово	а-
вдох	.хх
убыстрение темпа	> текст <
замедление темпа	< текст >
сомнительный отрезок	{текст}
неразборчивый отрезок	{--}
понижение громкости	* текст *
повышение громкости	ТЕКСТ
комментарии транскрибирующего	((текст))
смех сквозь сомкнутые губы	мхе

017-GornostayT:_Layout 1 18.05.2008 15:01 Page 109

СИСТЕМА МАШИННОГО ПЕРЕВОДА TILDE TRANSLATOR: НОВЫЙ ЭТАП В РАЗВИТИИ АНГЛО / РУССКО-ЛАТЫШСКОГО МАШИННОГО ПЕРЕВОДА

MACHINE TRANSLATION SYSTEM TILDE TRANSLATOR: NEW STAGE OF ENGLISH / RUSSIAN-LATVIAN MACHINE TRANSLATION

Горностай Т. *(tatjana.gornostaja@tilde.lv)*
компания Tilde, Рига (www.tilde.com)

В докладе представлен лингвистический программный продукт – система машинного перевода Tilde Translator. Внимание уделено особенностям разрешения лексической неоднозначности, переводу устойчивых словосочетаний и употреблению семантических фильтров для улучшения качества машинного перевода в рассматриваемой системе.

1. Введение. Машинный перевод – роскошь или необходимость?

Несмотря на общепризнанный факт о невозможности полностью автоматизировать и алгоритмизировать перевод текста с одного естественного языка на другой [1, 3, 5, 7, 9, 10, 11], машинный перевод (МП) представляет собой перспективное научное и практическое направление с почти 200 действующих коммерческих систем [12], в частности систем, производящих перевод текстов на латышский язык: SmartTranslator, Pragma Lite, Tilde Translator. Однако даже сегодня нередко можно услышать вопрос о том, для чего нужны системы МП, и человек в современном информационном пространстве часто оказывается в ситуации, когда ему неизвестны возможности применения языковых технологий для решения различных задач. Задачей как научных исследователей, так и разработчиков систем МП является ознакомление пользователя любой категории с возможностями той или иной системы МП для продуктивного ее использования [8].

Ни одна из систем МП не способна удовлетворить интересы всех пользователей. Следуя современной тенденции в теории МП [14, 15, 16], все системы МП можно разделить на две большие группы: разработки, ориентированные на технологии МП, способные максимально облегчить работу переводчика-профессионала, и на создание программ, обеспечивающих межъязыковое общение для обычного пользователя. Система МП Tilde Translator относится ко второму типу систем. Рассмотрим подробнее особенности данной системы МП.

2. Система машинного перевода Tilde Translator

В рамках традиционной типологии [2, 4, 17], лингвистическая информационная система Tilde Translator может быть классифицирована как многоязычная, полностью автоматизированная система МП. Входной единицей перевода может являться как отдельное слово языка-источника, так и словосочетание или предложение, а выходной единицей – соответствующая структурная единица выходного языка. В первом случае система будет работать как резидентный электронный словарь, извлекая из словарной статьи перевод иностранного слова, а получая на входе единицу, большую, чем отдельное слово, – как собственно система МП.

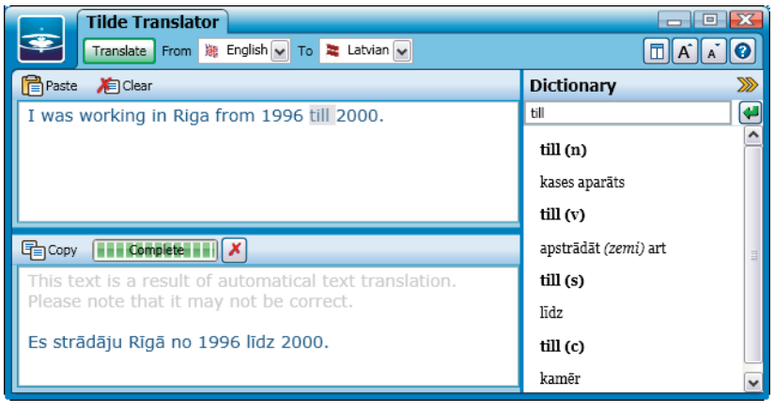


Рис. 1 Электронный словарь в системе МП Tilde Translator

017-GornostayT:_Layout 1 18.05.2008 15:01 Page 110

Надо сказать, что использование большинства систем МП в качестве электронного словаря не представляется рациональным, так как система МП, как правило, предложит лишь один вариант перевода заданного слова, а для более полной информации об искомой словарной единице следует обратиться к специально созданным для данных целей электронным словарям. В программе Tilde Translator реализована возможность вызова электронного словаря, интегрированного в систему МП для удобства пользователя в том случае, если система МП предлагает некорректный перевод того или иного слова, или пользователю необходимо просмотреть всю словарную информацию о данном слове (рис. 1).

2.1. Архитектура системы

Сегодня принято считать, что нет единственного метода для достижения МП необходимого качества, и будущее за моделями так называемого ‘гибридного’ МП, объединяющего лучшие технологии, основанные на правилах, статистике и переводческой памяти [15, 16]. Система МП Tilde Translator создавалась на базе современных тенденций в теории и практике МП и является одним из примеров подобных технологий МП, в основу работы отдельных модулей которой заложены статистические методы (модуль разрешения лексической неоднозначности, например) и технология переводческой памяти (перевод устойчивых словосочетаний) наряду со стратегией трансфера, составляющей ядро системы..

Архитектура рассматриваемой системы МП основывается на том же принципе, что и архитектура электронного словаря с элементами МП SmartTranslator [6, 13], – принципе модульной архитектуры. Алгоритмы перевода представляют собой многоуровневый процесс, в самом общем виде состоящий из трех этапов: анализ структуры входной единицы языка-источника; преобразование данной структуры в аналогичную структуру выходного языка (собственно трансфер); синтез единицы выходного языка в соответствии с полученной структурой на втором этапе [18].

2.2. Разрешение лексической неоднозначности

В нашем предыдущем докладе [6] мы рассматривали применение статистических методов в электронном словаре с элементами МП SmartTranslator для разрешения лексической неоднозначности в процессе перевода на латышский и литовский языки, при котором учитывалась вероятностная синтаксическая модель данных языков, построенная на основе одноязычного корпуса. Недостатком упомянутой разработки являлось использование данных выходного языка и игнорирование языка-источника. Была поставлена задача усовершенствования алгоритма разрешения многозначности при МП, и за последний год данный модуль претерпел изменения благодаря новой стратегии, которая на данный момент реализована для англо-латышского МП. При описанном подходе используется двуязычный англо-латышский параллельный корпус, на котором проводилось обучение системы, и инструмент GIZA++ для сопоставления переводных эквивалентов (выравнивания корпуса). Модуль разрешения многозначности подключается на том этапе МП, когда уже построены синтаксические деревья для языка-источника и выходного языка, и приписаны переводные эквиваленты для лексических единиц языка-источника. Итак, мы имеем один или несколько переводных эквивалентов в выходном языке, соотнесенных с одной вершиной дерева языка-источника. Для всех переводных эквивалентов в выходном языке вычисляется их вероятность перевода и выбирается переводной эквивалент с максимальным значением, учитывая локальный контекст данной лексической единицы в тренировочном корпусе. В результате, мы имеем возможность разрешения лексической неоднозначности, выбирая тот переводной эквивалент, который обеспечивает наивысшую вероятность перевода данной лексической единицы языка-источника.

Известные недостатки и неточности данного метода на сегодняшний день: ошибки при анализе корпуса и оценке перевода (морфологическая неоднозначность, например); игнорирование некоторых единиц в процессе анализа корпуса (числительные, например); использование переводных эквивалентов только из автоматического словаря системы (возможно, что в словаре вообще нет перевода для данной лексической единицы, или же в корпусе содержится более подходящий вариант).

2.3. Перевод устойчивых словосочетаний

Система МП, в отличие от систем переводческой памяти, предлагающей перевод ‘по образцу’, позволяет получать синтаксически связный перевод текста, учитывающий морфологию как языка-источника, так и выходного языка [4]. Своего рода модификацией технологии переводческой памяти в рассматриваемой системе МП Tilde Translator является усовершенствованный модуль перевода устойчивых словосочетаний, который

017-GornostayT:_Layout 1 18.05.2008 15:01 Page 111

производит лингвистический синтез устойчивых словосочетаний с учетом морфологии выходного языка и используется в англо / русско-латышском МП.

С точки зрения МП, устойчивыми единицами являются как идиоматические выражения, фразеологизмы, терминологические единицы, так и составные союзы и предлоги, которые должны обрабатываться как единое целое. Причем ‘многокомпонентность’ устойчивого словосочетания в рассматриваемой системе является общим понятием для пары языков и не обязательным критерием для каждого языка, то есть, либо в языке-источнике, либо в выходном языке устойчивое словосочетание может быть выражено одним словом, например:

(1) авторское право – autortiesības,
(2) автозавод – automobiļu rūpnīca,
(3) to be late – nokavēties,
(4) fire and fury – liesmainums.

Основное изменение в алгоритме перевода устойчивых словосочетаний – возможность обработки компонентов внутри структуры словосочетания. Напомним, что в предыдущей разработке корректный перевод получали лишь фиксированные, с грамматической точки зрения, устойчивые словосочетания [6].

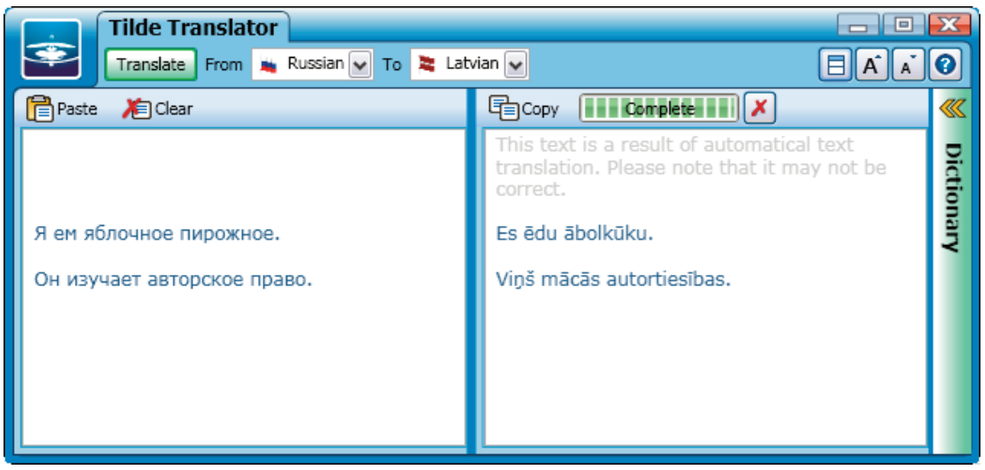


Рис. 2 Пример русско-латышского перевода устойчивых словосочетаний

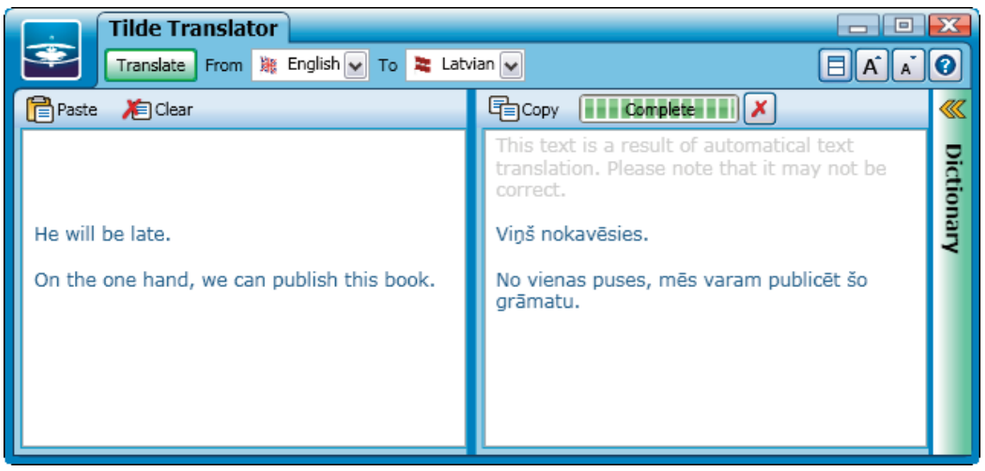


Рис. 3 Пример англо-латышского перевода устойчивых словосочетаний

Модуль обработки устойчивых словосочетаний представляет собой отдельный словарь, в котором хранятся устойчивые словарные единицы[1], а также набор разработанных правил трансфера, которые описывают, каким образом будет изменено синтаксическое дерево языка-источника при синтезе предложения выходного языка в случае несовпадения синтаксической структуры словарной единицы в языке-источнике и выходном языке. Проанализированное и переведенное устойчивое словосочетание интегрируется в синтаксическое дерево

[1] Около 25 тыс. словарных единиц для англо-латышского перевода, более 4 тыс. словарных единиц для русско-латышского перевода (в процессе пополнения).

017-GornostayT:_Layout 1 18.05.2008 15:01 Page 112

всего предложения, которое посылается для дальнейшей обработки (трансфер, согласование и другие процессы). На рисунках 2 и 3 приведены примеры русско-латышского и англо-латышского перевода устойчивых словосочетаний.

2.4. Семантические фильтры

Одним из решений в системе МП Tilde Translator является также использование семантических фильтров для лексических единиц языка-источника там, где это необходимо для улучшения качества МП. Лексические единицы могут быть отнесены к той или иной семантической категории. Категории не являются уникальными для данного направления и могут использоваться при переводе на другой язык.

Рассмотрим процесс установления семантического фильтра на примере синонимического ряда *say, tell, talk, speak* в английском языке. Принадлежность к одному синонимическому ряду не означает то, что все данные лексические единицы будут отнесены к одной семантической категории, так как мы используем семантические фильтры для описания структурных типов и конструкций, и не всегда лексические единицы с синонимичным значением будут образовывать однотипные синтаксические конструкции. Так, глаголы *talk* и *speak* отнесены к категории ‘communication’ на основе структурного сходства синтаксических конструкций, в которых они употребляются:
может употребляться без объекта (пример 5),
употребляются с предлогами *about* и *of* (пример 6, 7),
употребляются с предлогом *to* (пример 8).

(5) to be learning to talk / speak,
(6) to talk about / of something,
(7) to speak about / of something,
(8) to talk / speak to somebody.

Глаголы *say* и *tell* объединены в категорию ‘speech_declarative’ на основании следующих особенностей:
употребляются в предложениях с прямой речью (пример 9),
употребляются в предложениях с косвенной речью (пример 10, 11),
употребляются в конструкциях с пассивным залогом (пример 12, 13).

(9) “Hello,” somebody said / told,
(10) to say that / wh- word,
(11) to tell that / wh- word,
(12) something / somebody is said,
(13) something / somebody is told.

Рассмотрим примеры использования семантических фильтров при МП .

2.4.1. Применение семантических фильтров

Семантические фильтры используются на стадиях трансфера и согласования. В синтаксис трансформационных правил и правил согласования введен атрибут *KindOf* для установления принадлежности вершины дерева к семантической категории. На рисунке 4 приведен пример трансформационного правила, которое будет исполнено, при условии принадлежности имени существительного к семантической категории ‘time’.

```
TransferRule(PREP<-pcomp-N)  //pirms gada
{
        Parent.SourceSpelling  == "pirms";
        Child.KindOf  == "time";
        Parent.TargetSpelling  = "назад";
        Parent.POS  = ADV;
        Child.Case  = accusative;
        move_to_left(Child,Parent);
        MakeLink(Child-pcomp->Parent);
}
```

Рис. 4 Пример трансформационного правила с использованием семантического фильтра

017-GornostayT:_Layout 1 18.05.2008 15:01 Page 113

При переводе предлогов, например, учитывается информация о принадлежности лексической единицы, употребленной с предлогом, к той или иной семантической категории. Английский предлог *to*, например, употребленный с глаголом с семантическим фильтром ‘motion’ переводится на латышский язык предлогом *uz* (примеры 14, 15), а с семантическим фильтром ‘communication’ – предлогом *ar* (примеры 16, 17):

(14) to go to work – iet uz darbu,

(15) to take to a town – ņemt uz pilsētu,

(16) to talk to John – runāt ar Džonu,

(17) to speak to him – runāt ar viņu.

Этот же предлог, употребленный с глаголом с семантическим фильтром ‘speech_declarative’, не переводится на латышский язык, а косвенное дополнение будет синтезировано в форме дательного падежа, например:

(18) to say to somebody – sacīt kādam.

Для перевода именных предложных конструкций также учитывается семантическая информация о лексических единицах. Так, например, английский предлог *about* с существительным, принадлежащим к семантической категории ‘time’ или ‘period’, переводится на латышский язык предлогом *aptuveni*, в остальных же случаях – предлогом *par* (пример 19, 20):

(19) to work about five hours – strādāt aptuveni piecas stundas,

(20) to speak about something – runāt par kaut ko.

Семантические фильтры используются также для латышско-русского МП. Так, например, латышский предлог *pirms* может переводиться на русский язык предлогом *до* или наречием *назад* (примеры 21, 22), а предлог *pēc* – предлогами *через* и *после* (примеры 23, 24):

(21) pirms lekcijas – до лекции,

(22) pirms desmit gadiem – десять лет назад,

(23) pēc diviem gadiem – через два года,

(24) pēc darba – после работы.

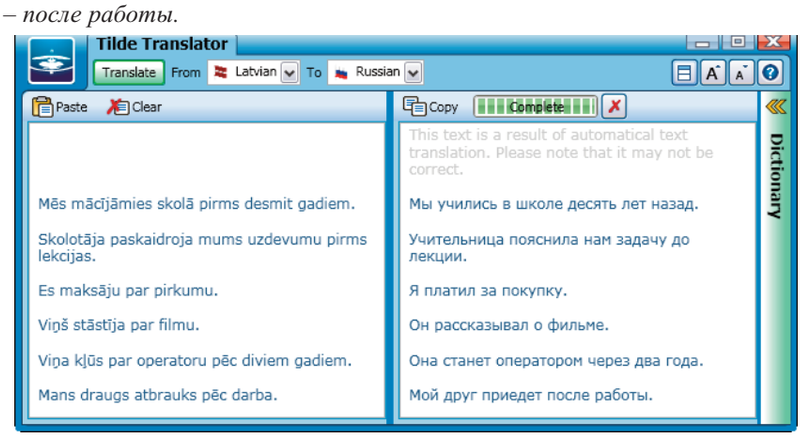


Рис. 5 Пример латышско-русского перевода с использованием семантических фильтров.

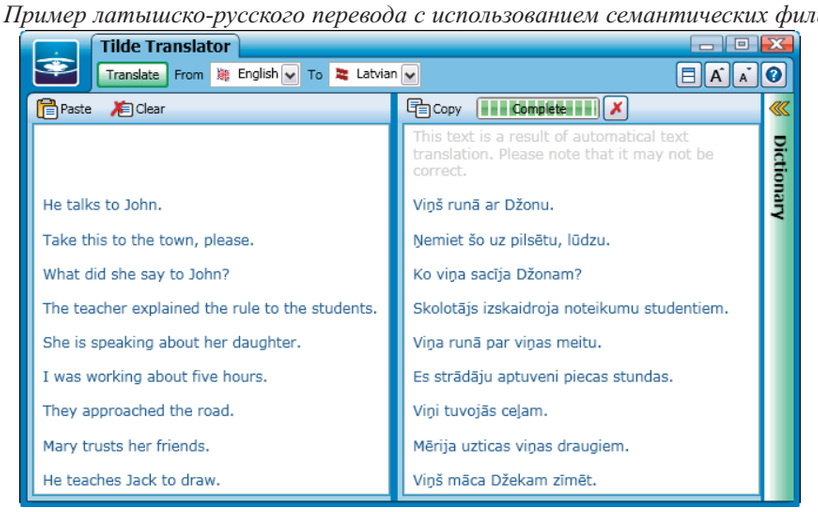


Рис. 6 Пример англо-латышского перевода с использованием семантических фильтров.

017-GornostayT:_Layout 1 18.05.2008 15:01 Page 114

Аналогично семантические фильтры используются также и для корректного перевода беспредложных глагольных конструкций, учитывая валентную структуру глагола в выходном языке (примеры 25, 26, 27):

(25) to approach the road – tuvoties ceļam,

(26) to trust somebody – uzticēties kādam,

(27) to teach somebody – to mācīt kādam.

Наряду со статистическим методом разрешения лексической неоднозначности используются и семантические фильтры. Приведем классический пример из латышско-русского перевода с использованием семантических фильтров 'publication' и 'food_plant' (примеры 28, 29). Рисунки 5 и 6 демонстрируют примеры латышско-русского и англо-латышского МП с использованием семантических фильтров.

(28) lasīt grāmatu – читать книгу,

(29) lasīt ogas – собирать ягоды.

4. Результаты и планы

Для оценки результатов МП системы Tilde Translator с использованием BLEU и NIST метрик автоматического измерения качества МП, полученного после внесения описанных изменений в систему, были составлены тестовые корпусы. Под 'тестовым корпусом' мы понимаем параллельный (выровненный по предложениям), не аннотированный (морфологически, синтаксически и лексически не маркированный), представительный и сбалансированный корпус текстов, который состоит из оригинальных предложений (500 единиц), то есть, корпус является естественным и достаточным для оценки качества МП данной системы.

Результаты для латышско-русского и англо-латышского направлений перевода обобщены в таблице 1. За базовый уровень взяты результаты оценки первой версии системы без введения семантических фильтров и нового подхода к разрешению лексической неоднозначности, а также переводу устойчивых словосочетаний.

Языковая пара	ЛАТ-РУС		АНГЛ-ЛАТ	
Метрика автоматического измерения качества МП	BLEU	NIST	BLEU	NIST
Базовый уровень	38,79	6,2762	10,87	3,2449
Последнее измерение (февраль 2008)	43,30	7,1857	21,15	5,3382

Таблица 1. Результаты автоматического измерения качества латышско-русского, англо-латышского МП

Полученные результаты позволяют говорить о положительных изменениях в системе с применением других технологий МП вместе со стратегией трансфера. Для русско-латышского направления автоматическая оценка качества МП проводилась на последней версии системы, полученные показатели: BLUE 17,98, NIST 4,22.

На будущее запланирована разработка модулей разрешения неоднозначности и перевода устойчивых словосочетаний для остальных направлений перевода, а также расширение автоматических словарей.

Автор доклада благодарит коллег Ингуну Скадиня, Дайгу Дексне и Карлиса Гоба за ценные замечания, высказанные в ходе написания статьи.

Список литературы

1. Беляева Л. Н. Проектирование лингвистических информационных баз для систем автоматической переработки текста // Статистика речи и автоматическая переработка текстов : межвуз . сб. науч. тр. / науч. ред. Р. Г. Пиотровский. – Л. : Ленингр. гос. пед. ин-т им. А. И. Герцена, 1988. – С. 8–28.

2. Беляева Л. Н. Автоматический (машинный) перевод / Л. Н. Беляева, М. И. Откупщикова // Прикладное языкознание : учебник / отв. ред. А. С. Герд ; С.-Петерб. гос. ун-т. – СПб. : Изд-во С.-Петерб. ун-та, 1996. – С. 360–388.

3. Беляева Л. Н. Лингвистические автоматы / Л. Н. Беляева, Р. Г. Пиотровский, В. Р. Нымм // Вестн. С.-Петерб. отд-ния Рос. акад. естеств. наук. – 1999. – [Т.] 3, № 1. – С. 73–82.

4. Беляева Л. Н. Лингвистические автоматы в современных информационных технологиях / Л. Н. Беляева ; Рос. гос. пед. ун-т им. А. И. Герцена. – СПб. : Изд-во Рос. гос. пед. ун-та им. А. И. Герцена, 2001. – 130 с.

5. Богуславский И. М., Иомдин Л. Л., Лазурский А. В., Митюшин Л. Г., Бердический А. С. Интерактивное разрешение неоднозначности различных типов в машинном переводе [Электронный ресурс] // Труды Междунар. конф. «Диалог' 2005»: [сайт] / Ассоц. компьютер. лингвистики и интеллектуал. технологий. – М., 2005. – Режим доступа:
http://www.dialog-21.ru/Archive/2005/Iomdin%20Boguslavski%20Lazurski/Iomdin%20Boguslavski%20Lazurski.htm.

017-GornostayT:_Layout 1 18.05.2008 15:01 Page 115

6. Горностай Т., Васильев А., Скадиня И., Скадиньш Р. Опыт латышско↔русского машинного перевода // ред. Л. Л. Иомдин, Н. И. Лауфер, А. С. Нариньяни, В. П.. Селегей. Труды Международной конференции Диалог' 2007 : Компьютерная лингвистика и интеллектуальные технологии. – М. : Издательский центр РГГУ, 2007. – С. 137-146.

7. Зубова И. И. Информационные технологии в лингвистике : Учеб. пособие / И. И. Зубова. – Минск : Изд-во Мин. гос. лингвист. ун-та, 2001. – 211 с.

8. Лингвистические ресурсы автоматизированного рабочего места филолога : Коллективная монография / Л. Н. Беляева, А. А. Виландеберк, Л. А. Деверь [и др.]. – СПб. : Изд-во ИнфоДа, 2004. – 184 с.

9. Пиотровский Р. Г. Инженерная лингвистика и теория языка / Р. Г. Пиотровский ; Акад. наук СССР, Науч. совет по теории сов. языкознания, Ин-т языкознания. – Л. : Наука, Ленингр. отд-ние, 1979. – 112 с.

10. Пиотровский Р. Г. Лингвистический автомат : (в исследовании и непрерыв. обучении) : Учеб. Пособие / Р. Г. Пиотровский ; Рос. гос. пед. ун-т им. А. И. Герцена. – СПб. : Изд-во Рос. гос. пед. ун-та им. А. И. Герцена, 1999. – 256 с.

11. Пиотровский Р. Г. Лингвистический автомат и его речемыслительное обоснование : учеб. Пособие / Р. Г. Пиотровский. – Минск : Изд-во Мин. гос. лингвист. ун-та, 1999. – 196 с.

12. Compendium of Translation Software [Electronic resource] : directory of commercial machine transl. systems and computer-aided transl. support tools / comp. by John Hutchins. – 14th ed. // EAMT : European Association for Machine Translation : [site]. – [Geneve], 2008. – Mode of access: http://www.eamt.org/compendium.html.

13. Deksne D., Skadiņa I., Skadiņš R., Vasiļjevs A. Foreign language reading tool – first step towards English-Latvian commercial machine translation system // Proceeding of the Second Baltic Conference on Human Language Technologies, Tallinn, 2005.

14. Hutchins W. J. Machine Translation and Human Translation : in Competition or in Complementation? // International Journal of Translation vol. 13, No 1-2, 2001, pp. 5-20.

15. Hutchins W. J. Towards a New Vision for MT [Electronic resource] : introductory speech… // Machine Translation Summit VIII, 18–22 Sept. 2001, Santiago de Compostela, Spain : papers and presentations / EAMT. – [Geneve], 2001. – Mode of access: http://www.eamt.org/summitVIII/papers/introduction.html.

16. Hutchins W. J. Current commercial machine translation systems and computer-based translation tools : system types and their uses // International Journal of Translation vol. 17, No 1-2, 2005, pp. 5-38.

17. Hutchins W. J. Machine translation : a concise history // To be published in Computer aided translation : Theory and practice, ed. Chan Sin Wai. Chinese University of Hong Kong, 2007.

18. Skadiņa I., Vasiļjevs A., Deksne D., Skadiņš R., Goldberga L. Comprehension Assistant for Languages of Baltic States // Nivre J., Kaalep H., Muischnek K., Koit M. (eds.) Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007. Tartu : University of Tartu, 2007, pp.167-174.

018-Graschenkova_A:_Layout 1 18.05.2008 14:36 Page 116

СТРУКТУРА ГРУППЫ ПРИЛАГАТЕЛЬНОГО В РУССКОМ ЯЗЫКЕ: ГИПОТЕЗА МАЛОГО a

RUSSIAN ADJECTIVE PHRASE: SPLIT AP HYPOTHESIS

Гращенкова А.Э. *(izmaja@mail.ru)*
Российский государственный гуманитарный университет

В данной работе предлагается анализ структуры группы прилагательного (ГП) в русском языке в рамках современной генеративной теории. На основании синтаксических и семантических свойств прилагательных делается вывод о сложной двухуровневой структуре ГП, включающей в себя внутреннюю лексическую проекцию AP и внешнюю функциональную оболочку aP.

Изучение синтаксических и семантических свойств прилагательных в рамках самых разных теоретических направлений никогда не пользовалось такой широкой популярностью как изучение глагольных и именных свойств. Тем не менее, находясь немного на периферии интересов различных теорий, исследования, посвященные прилагательным время от времени вызывали оживленные научные дискуссии. В рамках функционально-типологического парадигмы, подобный всплеск интереса был связан в первую очередь с фундаментальным *исследованием лексической типологии прилагательного, предложенного в статье Р.М.В. Диксона «Where have all the adjectives gone?»* (Dixon 1977), и продолженного в таких работах как (Givon 1979, Hopper&Thompson 1984, Croft 1991 и пр.). В рамках генеративного подхода интерес к изучению прилагательных был вызван в первую очередь исследованиями Д. Болинджера, Л. Бэбби, М. Сигел (Bolinger 1967, Babby 1971, 1973, Siegel 1976). Идеи, высказанные в этих работах, и до сих пор не перестают быть актуальными. Ниже мы остановимся именно на этой ветви исследований. В пункте 1 будет предложен небольшой исторический экскурс, имеющий целью ознакомить читателя с основными течениями исследований в рамках генеративизма, главным образом связанных с исследованиями прилагательных в русском языке. В п.2 будет предложен альтернативный синтаксический анализ структуры группы прилагательного в русском языке. В п.3 мы обсудим конкретные факты русского языка, представляющие, на первый взгляд, проблему для предложенного анализа.

1. Структура группы прилагательного: история вопроса

1.1. Краткая и полная формы

Исследователи прилагательного в русском языке сталкивались в первую очередь с проблемой дистрибуции кратких и полных форм прилагательных. Так, хорошо известно, что единственный синтаксический контекст, допускающий употребление кратких форм – первичная предикация:

(1.a) Мальчик (был) красив.
*(1.b) *красив мальчик*

В то же время полные формы допустимы как в собственно атрибутивном контексте, так и в предикативном:

(2.a) Мальчик (был) красивый.
(2.b) красивый мальчик

1.1.1. Полная форма: гипотеза нулевого имени (Null N Hypothesis)

Примеры типа (1)-(2) иллюстрируют несостоятельность подхода (общепринятого в традиционной грамматике), пытающегося объяснять синтаксические свойства полных и кратких форм исходя из их дистрибуции. Понимание этого факта естественным образом приводит к тому, что большинство исследований в рамках формальных теорий было направлено на поиск возможных объяснений наблюдаемой асимметрии в дистрибуции полных и кратких форм. А точнее, внимание исследователей главным образом сосредоточилось на поиске разумных объяснений допустимости предикативного использования полных форм прилагательных (пример (2.a)). И большинство анализов исходило из предположения, что данная асимметрия является лишь поверхностным проявлением глубинной симметрии, а именно, что в предикативном употреблении полные формы прилагательных всегда имеют скрытую именную вершину (нулевое N), а, следовательно, всегда атрибутивны.

018-Graschenkova_A:_Layout 1 18.05.2008 14:36 Page 117

В наиболее явной форме гипотеза нулевой вершины впервые была предложена в работах (Babby 1971, 1973, Siegel 1976). Основным и наиболее часто цитируемым аргументом данного анализа является контраст в интерпретации предложений типа:

(3.a) Студентка умна. — *абсолютная интерпретация*

(3.b) Студентка умная. — *относительная интерпретация*

В работах (Siegel 1976, Baylin 1992, Matushansky 2006, среди прочих) предлагается усматривать следующий (спорный на наш взгляд) семантический контраст между (3.a) и (3.b): предлагается считать, что первое из пары предложений несет в себе оценку субъекта (т.е. студентки) абсолютную, т.е. в данном случае не зависимую от умственных способностей других студенток; в то время как предложение (3.b) подразумевает оценку, базирующуюся на сравнении умственных способностей конкретной студентки с умственными способностями единиц входящих в множество $X[x_1 x_2 \dots x_n]$, такое что x – студентка. Ответственным за данную интерпретацию (3.b) считается нулевая вершина $Ø_N$, принадлежащая множеству X.

Таким образом, данный анализ предполагает следующую структуру для предложения (3.b):

(3'.b) Студентка умная $Ø_N$.

Подобный анализ, однако, не способен объяснить, например, следующих фактов русского языка:

♦ нарушение идиоматики:

(4) золотая рыбка *vs.* *рыбка (была) золотая*

(5) медные трубы *vs.* *трубы (были) медные*

Анализ в духе нулевого N не способен предсказать разрушение идиоматической связи, наблюдаемой в примерах (4)-(5).

♦ бессоюзное сочинение прилагательных:

(6) широкое синее море *vs.* *море широкое*$_{<\text{нисходящий тон, пауза}>}$ *синее*

(7) высокий каменный дом *vs.* *дом высокий*$_{<\text{нисходящий тон, пауза}>}$ *каменный*

В предикативной позиции бессоюзное сочинение прилагательных требует обязательного просодического маркирования (в частности, первый конъюнкт маркируется нисходящим тоном, с обязательной паузацией между двумя сочиненными прилагательными).

♦ запрет на лексические зависимые при полных формах в предикативной позиции:

(8.a) довольный своей оценкой студент

*(8.b) *студент (был) довольный своей оценкой*

Атрибутивные полные формы прилагательных могут иметь при себе лексические зависимые (8.a). Для предикативных полных форм такая возможность исключена (8.b).

Суммируя факты (4)-(8), отметим, что гипотеза, постулирующая единую структуру для полных форм в атрибутивной и предикативной позициях, не способна объяснить различия, тем не менее, демонстрируемые прилагательными в этих позициях.

1.1.2. Краткая форма

Л. Бэбби в (Babby 1971, 1973) предлагает трансформационный анализ кратких и полных форм (далее КФ и ПФ) прилагательных. Бэбби исходит из того, что главное отличие ПФ от КФ заключается в отсутствии признака падежа у последней, именно этим объясняется и запрет на атрибутивное использование КФ. Далее утверждается, что источником и для КФ и для ПФ является одна и та же глубинная категория глагола. При приобретении только признаков рода, лица и числа в поверхностной структуре мы получаем КФ (которая таким образом является обычным глаголом), если же добавляется ещё и категория падежа, то мы имеем дело с ИГ, а именно – с относительным предложением с нулевой вершиной (см. п. 1.1.1). Суммируя, КФ по Бэбби – глагол, обладает признаками и свойствами, характерными для «обычных» глаголов.

1.2. Общие вопросы, связанные со структурой адъективной группы

В этом разделе тезисно будут отмечены наиболее важные для дальнейшего изложения фрагменты анализов группы прилагательного (далее ГП) в разных языках:

♦ Р. Джакендофф в (Jackendoff 1977) выступает за расширение правил базовой структуры, предлагая считать, что прилагательное наряду с глаголом и именем возглавляет собственную фразовую категорию, развертывающуюся по X' схеме.

♦ В работах Г. Чинкве и Р. Кейна (Cinque 1994, Kayne 1994) формулируется гипотеза свернутого относительного предложения, в соответствии с которой ГП имеют структуру свернутого относительного предложения вида (9), субъектом (т.е., в терминах порождающей грамматики, спецификатором проекции IP) которого является определяемое имя:

018-Graschenkova_A:_Layout 1 18.05.2008 14:36 Page 118

(9) le $[_{FP}$ F^0 $[_{CP}$ $[_{XP}$ $jaune_j]$ $[C^0$ $[_{IP}$ [livre] $[I^0$ $[e]_j]]]]]$
le jaune livre
желтая книга

♦ Дж. Бернштайн в (Bernstein 1995) развивает идею свернутого относительного предложения[1]. Однако в предлагаемой структуре определяемое имя порождается как субъект AP (Adjective Phrase), которая в свою очередь порождается в позиции комплемента I (или другой функциональной вершины ниже I):

(10) $[_{DP}$ the $[_{CP}$ Spec $[C^0$ Ø $[_{IP}$ woman $[I^0$ Ø ... $[_{AP}$ Spec [proud [of her daughter]]]]]]]
the woman proud of her daughter
женщина, гордая своей дочерью

♦ Г. Чинкве (Cinque1990) и Х. Беннис (Bennis 1999,2000) на основе нескольких параметров (на материале итальянского, немецкого и голландского) выделяют две группы прилагательных: эргативные (ergative/unaccusative) и неэргативные (unergative). Суть различия заключается в способности прилагательного приписывать не падеж (как в случае неаккузативных глаголов), а тематическую роль (внутреннему или внешнему участнику, соответственно). Для неэргативных прилагательных в (Bennis 1999, 2000) постулируется наличие в структуре функциональной оболочки a:

(11) $[_{\underline{aP}}$ these people $[_{\underline{a}}$ $[_{AP}$ me $[_A$ loyal]]]
Эти люди преданы мне.

2. Структура ГП в русском языке:

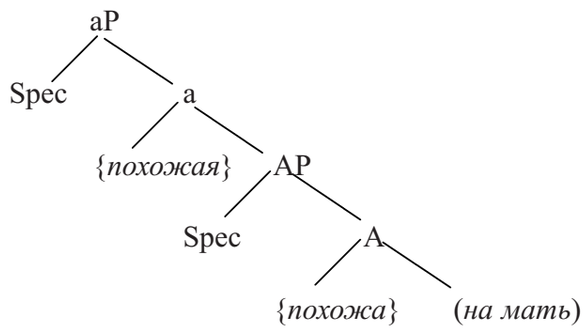
В работах (Гращенков, Гращенкова 2006, Grashchenkov&Grashchenkova 2007) отмечается следующий факт: в русском языке существует целый ряд прилагательных способных иметь собственные зависимые предложно-падежные группы (*полный, похожий, гордый, вежливый, довольный* и др.), такие прилагательные могут употребляться в ПФ в атрибутивной позиции (13), в КФ в предикативной позиции (12), однако запрещены в ПФ в предикативной позиции (14)[2]:

(12) Женщины были равнодушны к футболу.
(13) Лихорадка охватила даже равнодушных к футболу женщин.
*(14) *Женщины (были) равнодушные к футболу.*

Далее утверждается, что **в русском языке нет прилагательных с синтаксической валентностью, которые не располагали бы краткой формой**.

Основываясь на этих и некоторых других фактах, а также адаптируя идеи, высказанные в разное время в (Babby 1971,1973, Jackendoff 1977, Cinque1990, Bernstein 1995, Bennis 1999, 2000), в (Гращенков, Гращенкова 2006, Grashchenkov&Grashchenkova 2007) предлагается следующий анализ структуры ГП в русском языке:

(15)



(i) ГП, в максимально развернутом виде (как например в (13) и (15)), представляет собой двухуровневую структуру, включающую в себя вложенную лексическую проекцию AP и внешнюю функциональную оболочку aP;
(ii) AP поверхностно соответствует КФ прилагательного, aP соответствует полной форме (таким образом, когда в структуре присутствует только лексическая проекция, AP озвучивается как краткая форма прилагательного (*похожа на мать* в примере (15)); если структура представляет собой сложный двухуровневый комплекс вида $[_{aP}$ a^0 $[_{AP}$ A^0]], озвучивается полная форма прилагательного(*похожая на мать* в (15)));

[1] Дж. Бернштайн предлагает такую структуру не для всех прилагательных, а лишь для прилагательных с предложными зависимыми, а именно о таких прилагательных и пойдет речь ниже.
[2] Здесь мы намеренно упрощаем факты (ввиду ограниченного объема статьи), оставляя в стороне грамматичность предложений типа (i), с прилагательным в творительном падеже: (i) Он всегда был вежливым с соседями.

018-Graschenkova_A:_Layout 1 18.05.2008 14:36 Page 119

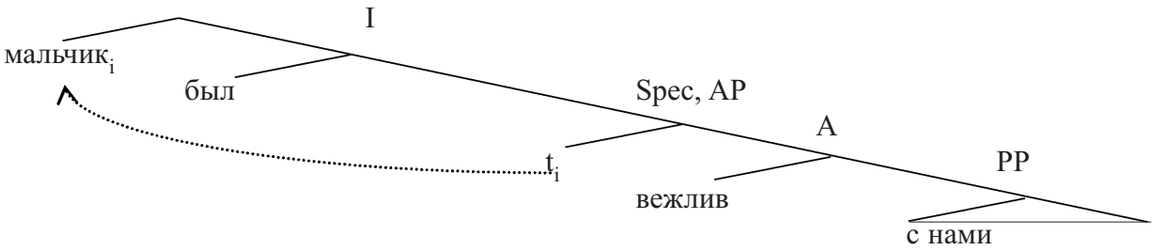
(iii) Только лексическая вершина А может приписывать семантические роли (таким образом, только наличие в структуре узла А имплицирует способность прилагательного иметь собственные зависимые);
(iv) Внешняя оболочка аР ответственна за падежное согласование – таким образом, ПФ возможны только при наличии в структуре внешней проекции аР;
(v) КФ прилагательных (то есть проекция АР) может быть напрямую выбрана из лексикона узлом I (тогда получается структура вида $[_{IP}\, I^0\, [_{AP}\, A^0]\,]$);

Рассмотрим несколько более подробно (не вдаваясь, однако в технические детали) процесс деривации полных и кратких форм, предполагаемый данным анализом[3].

2.1. Деривация кратких форм

(16) Мальчик был вежлив с нами.
(16.a) $[_{IP}$ *мальчик*$_i$ $[_I$ *был* $[_{AP}\, t_i$ $[_A$ *вежлив [с нами]]]]*

(16.b) Spec, IP



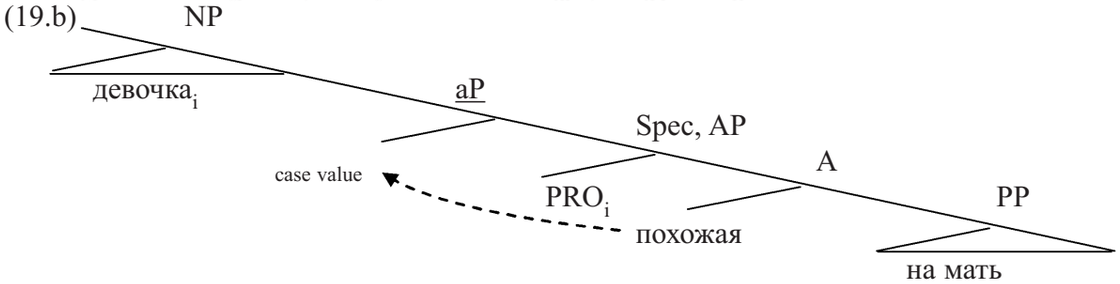
КФ прилагательного *вежливый*, в соответствии с (ii)-(iii) является вершиной проекции АР. В позиции спецификатора АР порождается и получает тематическую роль субъект предикации, ИГ *мальчик*, которая в дальнейшем подвергается передвижению в позицию Spec, IP (в соответствии с принципом развернутой проекции (Extended Projection Principle).

2.2 Деривация атрибутивных полных форм с зависимыми

ПФ прилагательных в атрибутивной позиции анализируются как правые адъюнкты к ИГ (см. похожий анализ в (Valois 1991, Bernstein 1993)). Правомерность такого анализа подтверждается существованием других правых адъюнктов в русском языке:
(17) мужчина [с черными усами]
(18) человек [из Кемерово]

Деривация ПФ с зависимыми в атрибутивной позиции (19.b) начинается так же, как и деривация КФ (см. (16.b)): прилагательное *похожий* начинает деривацию в позиции вершины АР (в соответствии с (ii)), далее следует передвижение прилагательного в вершину а, где происходит согласование по падежу с вершиной ИГ *девочка*.
(19) девочка, похожая на мать,…
(19.a) $[_{NP}$ *[девочка]* $[_{aP}$ *PRO*$_i$ $[_a$ *похожая]* $[_{AP}\, t_i$ $[_{A'P}$ t_H *[на мать]]]]*

(19.b)



[3] Ниже рассматривается деривация лишь тех структур, которые релевантны для дальнейшего обсуждения в п.3. Так, мы вынужденно (в силу ограниченного объема статьи) оставляем за рамками обсуждения деривацию предикативных полных форм в творительном падеже с зависимыми.

018-Graschenkova_A:_Layout 1 18.05.2008 14:36 Page 120

Важно отметить, что, в случае с ПФ, определяемое имя порождается слева от aP, в вершине N^0, и контролирует нулевое PRO, занимающее позицию субъекта (то есть спецификатора) AP. В пользу подобного анализа можно привести несколько аргументов.

Первый (положительный) аргумент связан с поведением анафоров в ГП. В (Гращенкова 2006), на основе поведения простого (*себя*) и составного (*самого себя*) рефлексивов делается вывод о необходимости постулирования нулевого PRO в структурах типа:

(20) $Я_i$ всегда мечтал найти человека$_j$,[$_{ГП}$ $Ø_j$ безжалостного ко всему на свете, в том числе и к себе $_{i, j}$ /самому себе$_{j/*\ i}$].

Наличие PRO в структуре (20) позволяет объяснить допустимость кореферентности субъектно-ориентированного[4] *себя* с определяемым именем в позиции прямого дополнения.

Теперь обратимся к негативному аргументу. Структуру в (19.b) можно было бы анализировать аналогично (16.b), как структуру, включающую в себя операцию подъема (raising) ИГ *девочка* из исходной позиции спецификатора AP. Однако если такой анализ кажется естественным для случая в (16.b), для (19.b) подобный анализ представляется проблематичным: во-первых, не понятно, что может вызвать передвижение ИГ из Spec AP в структуре типа (19.b) (напомним что в (16.b) передвижение вызывалось признаком EPP у узла I), во-вторых, кажется странной сама природа возможного передвижения XP в X^0 (в (16.b) происходило передвижение XP в XP).

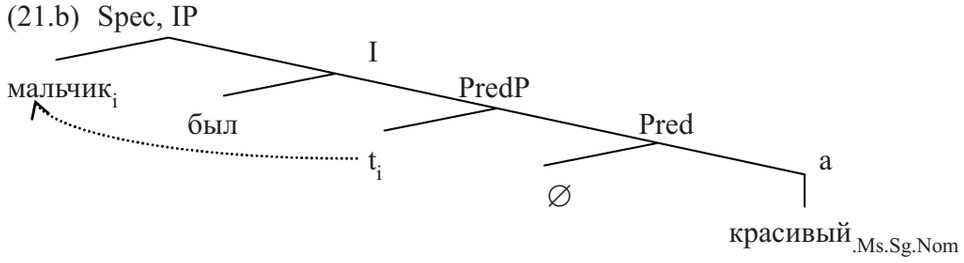
2.3. Деривация предикативных полных форм

Для ПФ одиночных прилагательных в предикативной позиции предлагается довольно стандартный анализ в духе (Bowers 1993, Baylin 2001). Прилагательное, не проецирующее внутренней структуры, является лексическим комплементом проекции PredP. В позиции спецификатора PredP порождается и получает тематическую роль субъект предикации, который в дальнейшем подвергается передвижению в Spec, IP (ср. (16.b)).

(21) Мальчик (был) красивый

(21.a) [$_{IP}$ $Он_i$ [$_I$ (был) [$_{PredP}$ t_i [Pred [$_a$ красивый]]]]]

(21.b)



Вслед за (Strigin & Demjjanow 2001) мы считаем, что именительный падеж (согласовательный) лексической вершины в именной предикации обусловлен согласовательной цепочкой I-Pred-Adj, где вершина Pred, лишенная собственных согласовательных признаков, является передатчиком (transmitter) падежного признака из I в лексическую вершину (в нашем случае вершину a). Ключевое условие передачи падежного признака заключается в том, что для того чтобы образовалась эта согласовательная цепочка, падежный признак на I должен быть активирован, то есть в Spec IP должна находиться ИГ с проверенным признаком падежа. Иными словами падежное согласование по цепочке I-Pred-Adj запускается в тот момент, когда заполняется позиция Spec, IP.

2.4. Деривация предикативных полных форм с зависимыми

Теперь посмотрим, как предлагаемый анализ объясняет неграмматичность ПФ с зависимыми в предикативной позиции, то есть неграмматичность примеров типа:

(22) *Мальчик (был) вежливый с нами.

В соответствии с (iii), наличие зависимых предполагает присутствие в структуре вершины A. Субъект предикации, в таком случае, так же как и в (16.b), должен порождаться в Spec, AP[5]. Для успешного завершения деривации прилагательное должно подвергнуться перемещению в вершину a, где происходит падежное согласование

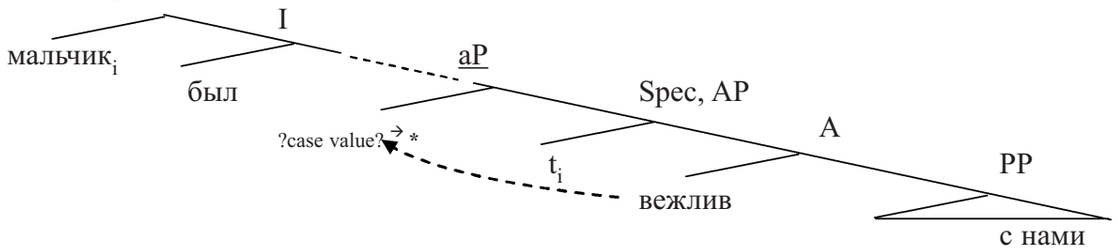
[4] О субъектной ориентированности, а также других свойствах рефлексива *себя* см. (Rappaport 1986, Падучева 1983, Тестелец &Толдова 1998, Лютикова 2002).

018-Graschenkova_A:_Layout 1 18.05.2008 14:36 Page 121

(а точнее, копирование падежа субъектной ИГ). Так же как и в структуре (21.b), падеж на вершине а может быть лишь результатом трансмиссии из вершины I. Однако как мы помним из п.2.3., трансмиссия может быть запущена лишь после того, как будет заполнена позиция Spec, IP. Но ИГ мальчик в (22.b) не может быть выдвинута из аР (и таким образом не может попасть в позицию Spec,IP) до тех пор, пока не завершена деривация самой аР, то есть до тех пор, пока не произойдет перемещение прилагательного из A^0 в a^0 и т.д. Деривация не может быть завершена успешно и предложение (22) – неграмматично.

(22.a) *$[_{IP}$ *Мальчик* $_i$ $[_I$ *был* $[_{\underline{aP}}[_a$*вежливый*$_k[_{AP}$ t_i $[_A$ t_k *[с нами]]]]]]]]*

(22.b) Spec, IP



Итак, анализ структуры ГП в русском языке, предложенный в (Гращенков, Гращенкова 2006, Grashchenkov&Grashchenkova 2007) справляется с поставленной задачей объяснения грамматичности предложений типа (12)-(13) и недопустимости примеров типа (14). Тем не менее, данный анализ на первый взгляд противоречит некоторым другим фактам, к изложению которых мы приступаем в п.3.

3. Проблемные факты

3.1. ДЛЯ - зависимые

→ Проблема

Все гипотезы о структуре прилагательных с зависимыми в п.2 строились на предположении, что в русском языке нет прилагательных с синтаксической валентностью, которые не располагали бы краткой формой.

Однако более тщательный анализ материала позволяет выявить целый ряд прилагательных, не имеющих краткой формы, но способных присоединять зависимые предложные группы (*ранний, поздний, взрослый, общий, главный, ключевой, драматический, победный, дебютный, центральный*).

Оказывается, что такие прилагательные могут иметь только один тип зависимых, а именно – предложные группы, возглавляемые предлогом *для* (далее *для*РР):

(23) Несмотря на еще ранний для Москвы час, на улицах народу было много, и видно — не занятого делами.[6]

(24) Ему хотелось есть; но что же мог он добыть в такой поздний для Кладенца час?

(25) Он предполагает снижение ключевых для экономики налогов.

(26) В целом вы довольны качеством борьбы, которое показали на дебютном для себя чемпионате мира?

(27) Последний день четвертьфиналов чемпионата мира по футболу стал драматическим для хозяев.

→ Решение

Возможный анализ: *для*РР в структуре адъективной группы занимают позицию адъюнкта при верхней составляющей (аР).

Подобный анализ подтверждается следующими фактами:

Прилагательные, типа *близкий* и *крупный* располагают краткой формой, в атрибутивной функции могут иметь *для*РР зависимые, <u>в краткой форме *для*РР зависимые запрещены:</u>

*(28) *Он (был) близок для меня.*

*(29) * Цветок (был) неожиданно крупен для таких веточек.*

Ср. допустимость *для*РР с полными формами этих же прилагательных:

(30) Он не позволял по отношению к себе́ никакого амикошонства, если дело касалось «посторонних», то есть людей, не принадлежащих к самому близкому для него́ кружку́, то есть «всех нас».

(31) Сам же цветок неожиданно крупный для таких субтильных «веточек» — о? коло 5 см диаметром, и яркий— жёлтый с красным зевом.

---

[5] Субъект не может начать деривацию в Spec IP, поскольку сначала должен получить тематическую роль.

[6] Все примеры, используемые в п.3, взяты из Национального Корпуса Русского Языка.

018-Graschenkova_A:_Layout 1 18.05.2008 14:36 Page 122

? Контрпримеры

Существует ряд прилагательных позволяющих *для*РР зависимые как в атрибутивной, так и в предикативной позициях: *вредный, полезный, важный* и пр.:

(32) Китайские врачи считают, что еда с кислым вкусом очень полезна для печени и почек.

(33) Даже слабоалкогольные напитки типа пива вредны для здоровья.

(34) Поэтому его язык очень труден для восприятия.

Очевидным образом лексемы, допускающие *для*РР как в КФ, так и в ПФ образуют единый семантический тип прилагательных с бенефактивным участником. Этот участник синтаксически может быть также выражен дативной ИГ или предложной группой для + генитив. То есть *для*РР в случаях типа (32)-(33) является семантическим аргументом, а не адъюнктом как в (23)-(27), чем и объясняется грамматичность (32)-(34).

3.2. Скрытые комплементы

3.2.1. Симметричные прилагательные

Прилагательные типа *аналогичный, идентичный, параллельный, похожий, равный, тождественный, эквивалентный* и пр. допускают выражение двух или более участников подлежащим во множественном числе, либо сочиненным подлежащим.

Такие прилагательные не могут употребляться в полной форме в номинативе после связки:

*(35) *Петя и Вася (были) похожие.*

*(36) *Выражения X и Y (были) эквивалентные.*

Несмотря на наличие одной ИГ в поверхностной структуре, внутренняя структура симметричных прилагательных содержит двух участников:

(37a) $[_{AP}$ *X похожий Y* $]$

(37b) $[_{AP}$ *X+Y похожий* t_Y $]$

Таким образом, проблема отсутствия данной формы у симметричных прилагательных сводится к общей проблеме отсутствия предикативного номинатива у прилагательных с зависимыми (см. структуру (22.b)).

НО: прилагательное *одинаковый*

Прилагательное *одинаковый*, являющееся на первый взгляд представителем класса симметричных прилагательных, допускает предикативный номинатив полной формы:

(38) Выбрать из них одну́ было трудно, потому́ что близнецы́ были почти́ одинаковые.

(39) Все мы здесь одинаковые!

Другое существенное отличие прилагательного о*динаковый* от класса симметричных прилагательных заключается в том, что оно не может присоединять лексических зависимых:

*(40) *Наш дом (был) одинаков с домом Петровых.*

*(41) * … наш дом одинаковый с домом Петровых…*

Следовательно, можно предположить, что множественный субъект (как в (38)-(39)) порождается не внутри проекции AP, а выше (например, Spec,PredP) как и в случае предикативных полных форм прилагательных без зависимых (см. структуру (21.b)).

3.2.2. Несимметричные прилагательные

Запрет на предикативный номинатив действует и в случае некоторых «обычных» прилагательных:

*(42) *Мальчик был благодарный.*

*(43) *Игрок был необходимый.*

Неграмматичность примеров (42)-(43) естественным образом вызывает предположение, что мы имеем дело с ситуацией сохранения внутренней структуры (то есть внутренней проекции AP) при отсутствии выраженных зависимых (как и в случае с симметричными прилагательными). То есть неграмматичность (42)-(43) является результатом более общего запрета на предикативный номинатив полных форм с зависимыми (см. структуру (22.b)).

? НО: существуют прилагательные типа *вежливый*, способные присоединять зависимые, однако не демонстрирующие запрета на предикативный номинатив:

(44) Мальчик был вежливый.

(45) Мальчик был вежливым с соседями.

? Почему

В п. 3.2.2. мы столкнулись с фактами (неграмматичность (42)-(43) и приемлемость (44)), необъяснимыми на первый взгляд в терминах анализа, предложенного в п. 2. Однако прежде чем отказываться от теории, спо-

018-Graschenkova_A:_Layout 1 18.05.2008 14:36 Page 123

собной объяснить многие другие факты, кажется целесообразным несколько более пристально взглянуть на факты ей противоречащие. Для этого обратимся к семантике конструкций с ПФ одиночных прилагательных в предикативной позиции.

2.2.1. Stage vs. Individual level[7] интерпретация

Рассмотрим примеры:

(46) Девушка (была) доступная

*(47) *Информация (была) доступная*

Примеры (46)-(47) демонстрируют ещё один удивительный факт: и в том, и в другом предложении лексической вершиной именной предикации является прилагательное *доступный.* Тем не менее (46) грамматично, а (47) – нет.

Однако с точки зрения семантики именного предиката (в терминах (Carlson 1977, Kratzer 1995) в (46) и (47) мы имеем дело с разными семантическими интерпретациями предиката *доступный*. Так, в (46) *доступный* имеет значение внутреннего качества (individual level). В (47) объекту приписывается состояние, релевантное в течение ограниченного отрезка времени (stage level). Это наблюдение позволяет сделать следующий промежуточный вывод:

(Не)возможность предикативного номинатива связана с интерпретацией.

В примерах (44)-(45) в случае с *вежливый* наблюдается та же корреляция между значением прилагательного и типом синтаксической конструкции. (44)- individual level, (45) – stage level.

Более того, то же верно и для случая с *благодарный:* неграмматичность (42) связана со stage-level интерпретацией прилагательного. Ср. допустимость идиоматизированного:

(48) Публика была благодарная.

В (48) мы, безусловно имеем дело с individual level интерпретацией, и как следствие с отсутствием запрета на предикативный номинатив.

Наблюдения над семантическими свойствами ПФ прилагательных в предикативной позиции позволяют сделать следующее утверждение:

♦ Individual level интерпретация – возможность предикативного номинатива

♦ Stage level интерпретация – запрет на предикативный номинатив

Вспомним, что анализ представленный в п.2 постулирует запрет на предикативный номинатив адъективных групп с внутренней проекцией АР.

Таким образом, возникает законный вопрос – насколько стадиальная интерпретация связана с наличием у прилагательного внутренней структуры? В работах (Stowell 1983, Kratzer 1995) на материале английского языка было показано, что наличие внутренней структуры (зависимых групп) имплицирует стадиальную интерпретацию.

Суммируя вышеизложенные наблюдения, заключаем: за разные интерпретации отвечают разные структуры. Предложение (48) *Публика была благодарная* (individual level) имеет структуру:

(49) $[_{IP}$ *Публика*$_i$ $[_I$*(была)* $[_{PredP}$ t_i *[Pred* $[_{\underline{a}}$ *благодарная]]]]] (т.е как в п.2.3)*

Предложение (42) **Мальчик благодарный* (stage level) имеет структуру:

*(50) **$[_{IP}$ *Мальчик* $_i$ $[_I$ *был* $[_{\underline{aP}}[_a$*благодарный*$_k$ $[_{AP}$ t_i $[_A$ t_k *]]]]]] (т.е как в п.2.4)*

Мы начали работу, предложив гипотезу о расщепленной структуре ГП в русском языке. Изучение фактов, на первый взгляд не объяснимых в рамках данного анализа, позволило нам сделать несколько важных, по нашему мнению выводов: целый ряд прилагательных в русском языке проецирует сложную двухуровневую структуру; внутренняя лексическая проекция отвечает за присвоение семантических ролей и соответствует стадиальной интерпретации; внешняя функциональная оболочка лишена всех этих свойств и отвечает только за согласование по падежу.

[7] Термины *stage level* и *individual level* были впервые предложены и обоснованы в работе (Carlson 1977) и далее разрабатывались в работах А. Кратцер (см. Kratzer 1995). Термины stage level и individual level соответствуют принятым в традиционной русистике терминам состояние и качество, однако в отличие от последних определяются не интуитивно, а на основе синтаксических тестов.

018-Graschenkova_A:_Layout 1 18.05.2008 14:36 Page 124

Список литературы

1. Babby L. 1971. A Transformational Analysis of Russian Adjectives. Ph. D. Thesis, Harvard University, Cambridge.

2. Babby, Leonard H. 1973. "The Deep Structure of Adjectives and Participles in Russian", Language, 49.349-360

3. Baylin, J. (1994) «The Syntax and Semantics of Russian Long and Short Adjectives: an X'-Theoretic Account,» in J. Toman (ed.) Formal Approaches to Slavic Linguistics. The Ann Arbor Meeting. (pp. 1-30) Ann Arbor, MI: Michigan Slavic Publications.

4. Baylin, J. 2001. The Syntax of Slavic Predicate Case. In A. Strigin et al (eds.) ZAS Occasional Papers in Linguistics: Proceedings of the Workshop on Predication, Zentrum für allgemeine Sprachwissenschaft, Berlin.

5. Bennis, H. 2004. Unergative Adjectives and Psych Verbs. In: Alexiadou, Artemis, Elena Anagnostopoulou and Martin Everaert (eds.), The Unaccusativity Puzzle: Studies on the Syntax-Lexicon Interface, Oxford.

6. Bernstein, J. 1993. Topics in the Syntax of Nominal Structure Across Romance. CUNY Ph.D. Dissertation.

7. Bernstein, J. 1995. Adjectives and their complements. Paper presented at the 1995 LSA Annual Meeting.

8. Bolinger D. 1967. Adjectives in English: Attribution and Predication. Lingua 18: 1-34.

9. Bowers, J. 1993. The Syntax of Predication. Linguistic Inquiry 24, 591-656.

10. Cinque, G. 1990. Ergative Adjectives and the Lexical Hypothesis. Natural Language and Linguistic Theory 8: 1-40.

11. Cinque, G. 1994. "On the evidence for partial N movement in the Romance DP," in Cinque, G, J. Koster, J.-Y.

12. Dixon R.M.W. 1977. 'Where have all the adjectives gone?' // Studies in language 1: 19-80.

13. Givon T. 1979. On understanding grammar. New York Academic Press.

14. Grashchenkov P., Grashchenkova A 2007 Argument Structure of Russian Adjectives. Paper presented at Workshop on Argument Structure and Syntactic relations. Vitoria-Gasteiz 23-25 May 2007.

15. Hopper P.J., Thompson S.A. 1984. 'The discourse basis for lexical categories in universal grammar' // Language 60: 703-752.

16. Jackendoff, R. 1977. The status of thematic relations in linguistic theory. Linguistic Theory. 18 (3). 369-411.

17. Kayne, R. 1994. The Antisymmetry of Syntax. MIT Press, Cambridge (MA). Linguistics, Oxford University Press.

18. Kratzer, Angelika 1995. Stage-level and Individual-level Predicates. Gregory Carlson & Francis Jeffrey Pelletier (eds.) The Generic Book. The University of Chicago Press, 125-75.

19. Matushansky, O. 2006. How to be Short: Some Remarks on the Syntax of Russian Adjectives. Séminaire de l'UMR.

20. Rappaport G. 1986. 'On anafor binding in Russian' // Natural language and linguistic theory 4: 97-120.

21. Siegel, M. E. A. 1976. Capturing the Russian Adjective. In B. H. Partee, ed., Montague Grammar. New York: Academic Press, pp. 293-309.

22. Stowell, Tim. 1983. Subjects across categories. The Linguistic Review 2:285-312.

23. Strigin A. & Demjjanow A. 2001 Measure Instrumental in Russian. Papers on Predicative Constructions, Berlin.

24. Гращенков П.В., Гращенкова А.Э. 2006. Аргументная структура прилагательных. Доклад на Третьей Конференции по типологии и грамматике для молодых исследователей. С.-Петербург ИЛИ РАН.

25. Гращенкова 2006 – А.Э. Гращенкова. Рефлексивы в группе прилагательного: теоретические проблемы и типология. // Вопросы языкознания, №1, 2006.

26. Лютикова Е. А. 2002. Когнитивная типология: рефлексивы и интенсификаторы. М.: ИМЛИ РАН.

27. Падучева Е. В. 1983. Возвратное местоимение с косвенным антецедентом и семантика рефлексивности // Семиотика и информатика Вып. 21.

28. Тестелец Я. Г., Толдова С. Ю. 1998. Рефлексивные местоимения в дагестанских языках и типология рефлексива. // Вопросы языкознания, № 4, 1998.

019-Grishina_E:_Layout 1 18.05.2008 14:36 Page 125

КОРПУС ЗВУЧАЩЕЙ РУССКОЙ РЕЧИ В СОСТАВЕ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА. ПРОЕКТ*

CORPUS OF ORAL RUSSIAN IN THE FRAMEWORK OF RUSSIAN NATIONAL CORPUS. CONSTRUCTION PROJECT

Гришина Е.А. *(rudi2007@yandex.ru),* ***Савчук С.О.*** *(savsvetlana@mail.ru)*
Институт русского языка им. В.В.Виноградова РАН

Статья содержит описание проекта «Корпуса звучащей русской речи», который может быть создан на материале подкорпуса «Речь кино» в составе Национального корпуса русского языка. В статье предлагаются предварительные решения, касающиеся структуры корпуса, типов разметки, формата выдачи материала, типов пользовательских запросов и разновидностей задач, которые могут быть решены посредством данного корпуса.

1. Постановка проблемы

Как известно, Национальный корпус русского языка включает в себя подкорпус устной речи (ср. Гришина 2005, Grishina 2006, Савчук 2008). В данный момент этот подкорпус выведен из состава основного корпуса и функционирует как самостоятельный модуль в рамках НКРЯ (наряду с поэтическим, диалектологическим, параллельным, образовательным и акцентологическим (в недалеком будущем) подкорпусами). Устный подкорпус включает в себя расшифровки реальной устной речи (публичной и частной, общий объем 4,4 млн словоупотреблений), а также кинотранскрипты (объем 1,6 млн словоупотреблений)[1].

Естественно, с самого начала функционирования устного подкорпуса возникал вопрос – и у самих создателей НКРЯ, и у его пользователей – о возможности создании корпуса не просто устной, но звучащей речи. И в силу ряда причин, в основном субъективного, но также и объективного характера этот вопрос регулярно получал отрицательный ответ[2].

Данная статья рассматривает следующую проблему: предположим, было бы принято решение начать создание корпуса звучащей русской речи (КЗРР) в рамках НКРЯ, – какой в этом случае могла бы быть технология создания этого корпуса? Каковы те задачи, которые было бы разумно ставить перед собой на начальных этапах, в какой последовательности и какими средствами и способами их решать?[3]

2. Источники КЗРР

Очевидным источником КЗРР являются расшифровки устных текстов, уже включенные в состав НКРЯ, для которых имеются соответствующие звуковые файлы, – казалось бы, достаточно разбить на небольшие блоки

* Статья написана при поддержке грантов РФФИ 06-06-80133а и 08-06-00371а.

[1] Разработка корпусов устной речи ведется по двум направлениям, в рамках которых решаются различные научные и практические задачи, что обусловливает и существенные различия в принципах и технологии создания корпусов: 1) фонетические (акустические, речевые) базы данных; 2) корпуса устных текстов (подробнее см. Савчук 2007). В существующих корпусах устных текстов на русском языке используются варианты графической фиксации звучащей речи – либо специально разработанная дискурсивная транскрипция, как в корпусе детских рассказов о сновидениях (Кибрик, Подлесская 2003; Подлесская, Кибрик 2004), либо упрощенная фонетическая транскрипция на базе русской орфографии, как в диалектном корпусе в составе НКРЯ, либо орфографическая запись с передачей отдельных фонетических особенностей устного дискурса, как в подкорпусе устной речи НКРЯ.

[2] Отметим, что за рубежом подходы к созданию аналогичных корпусов уже осуществляются, в основном в рамках психолингвистических исследований эмоций, ср. Clavel et al 2006, Devillers et al. 2006, исследований детской речи, ср. широко известный проект CHILDES, а также в образовательных проектах, ср. Braun et al. 2007.

[3] Предлагаемые в статье решения, безусловно, носят субъективный и вполне дискуссионный характер, но практика работы над Национальным корпусом показывает, что максимальные результаты достигаются только при наличии личной заинтересованности исполнителя в проекте, а неизбежной платой за это является определенная субъективность принятых решений.

019-Grishina_E:_Layout 1 18.05.2008 14:36 Page 126

скрипты и звуковые файлы, выровнять их между собой, добавить метаразметку, морфологическую и семантическую разметку, и корпус звучащей речи готов.

Однако, по зрелом размышлении, это очевидное решение следует признать если не неверным, то недостаточным. Мы считаем, что начинать строительство КЗРР следует с кинотранскриптов, включенных в состав НКРЯ, и соответствующих фильмов[4]. Аргументы здесь таковы.

Во-первых, кино содержит, помимо звукового, зрительный ряд, что позволяет при создании КЗРР поставить вопрос об одновременном создании корпуса русских жестов – то есть одна и та же операция (нарезка мультимедийного файла на блоки и их выравнивание с текстом) в случае работы с кинотранскриптами дает пользователю в два раза больше возможностей получения информации, чем в случае работы с доступными в настоящий момент создателям НКРЯ аудиофайлами[5].

Во-вторых, никакая, сколь угодно большая коллекция аудиофайлов не даст такого разнообразия устной речи, как кинотранскрипты, – хотя бы потому, что кино в большом количестве включает в себя ситуации, в которых в реальной жизни никакая аудиозапись в нормальном случае невозможна, а если и возможна, то результаты такой аудиозаписи для нас недоступны[6].

Очевидным недостатком принятого решения является тот факт, что в КЗРР включается «ненатуральная», «неестественная» устная речь. Здесь трудно что-либо возразить, кроме, пожалуй, следующего. Ненатуральность киноречи – миф, если не полностью, то в значительной степени (см., в частности, работу Капанадзе 1986 об ориентации языка кино на реальную разговорную речь). Нам уже приходилось писать о том, что речь кино – это результат совершенно особого процесса усвоения чужого текста актером и превращения этого текста в его, актера, собственный текст (см. Grishina 2007). И можно даже утверждать, что успех фильма в значительной степени зависит от того, насколько удачно и полно пройдет этот процесс присвоения[7].

С другой стороны, естественное повседневное общение протекает в стереотипных формах, усвоенных и воспроизводимых в стереотипных ситуациях. Не случайно в научных описаниях речевого взаимодействия, активно используются театральные метафоры: «Повседневное, бытовое по преимуществу, общение (персональный дискурс) обслуживает *сценарии* социального взаимодействия, которые можно уподобить принципам *комедии dell'arte*, где при достаточно четкой определенности *характеров действующих лиц актерам* предоставляется значительная свобода в содержании *реплик*» (Седов 2007: 21)[8].

Безусловным «проигрышем» речи кино по сравнению с реальной устной речью является повышенная связность устного текста в кинематографе. Но здесь речь кино проигрывает в основном частной, приватной устной речи, в особенности бытовой, где разговаривают хорошо знакомые люди, для которых нет необходимости проговаривать все до конца. Если мы сравним киноречь с публичной устной речью, то здесь различия по связности нивелируются, а что касается естественности человеческого речевого поведения (обрыв, наложение реплик, перебивы, запинки, смены стратегий и проч.), то по этому параметру в ряде случаев киноречь даст фору некоторым образцам публичной устной речи.

Как бы то ни было, нам представляется, что плюсы в случае работы с кинотранскриптами существенно мощнее, чем минусы.

---

[4] Безусловно, это не значит, что в качестве второго эшелона скрипты реальных устных текстов не должны быть включены в КЗРР, – речь в данный момент идет именно о начальном этапе работы.

[5] Таким образом, проект нацелен на реализацию многократно высказывавшейся, начиная с пионерских работ Л.П. Якубинского (1923) и В.Н. Волошинова (1930), идеи о том, что устную речь необходимо изучать в связи с ситуацией общения, в которой она порождается, поскольку сама внеязыковая ситуация становится частью акта коммуникации, а в передаче смысла сообщения участвуют не только вербальные, но и невербальные средства – жесты, мимика, интонация и под.; лингвистические описания русской разговорной речи с опорой на ситуацию находим, в частности, в работах Земская и др. 1981, Китайгородская, Розанова 2005. Корпус звучащей речи в этом плане резко расширяет возможности исследователя. В последнее время аудио- и видеофиксация материалов с целью создания корпусов и баз данных начинает использоваться и в других областях филологии: при документировании малых языков (Кибрик А.Е. и др. 2006), в фольклористике (Кляус 2004, Мороз 2003).

[6] Именно богатство и разнообразие жизненных ситуаций, разыгранных в кинематографе, дает нам возможность максимально полно исчислить те параметры, которые следует предусмотреть для разметки звуковых файлов, – стандартные записи устной речи производятся обычно в слишком «тепличных» условиях. Таким образом, совокупность параметров, предусмотренных для описания киноклипов, заведомо будет включать в себя параметры, необходимые для описания реальной устной речи, но не наоборот.

[7] Следствием этого процесса присвоения чужого текста является совершенно однозначное поведение в речи кино т. н. маркеров устной речи (см. Гришина 2007) – по этому параметру речь кино и реальная устная речь практически не отличаются. Свидетельством того, что устная речь в кино не подвергалась и не подвергается никакой особенной редактуре, служит существенное число ошибок, встречающихся в киноречи, – лексических, фактических и стилистических.

[8] Ср. также широко распространенные терминологические сочетания: роли говорящих, речевой сценарий, репертуар речевых жанров и пр.

019-Grishina_E:_Layout 1 18.05.2008 14:36 Page 127

3. Структура КЗРР

3.1. Единица выдачи и основные типы запросов

Предполагается, что единицей выдачи в КЗРР являются объекты двух типов:

1) морфологически и семантический размеченный текст (в соответствии с параметрами и методикой, принятыми при разметке текстов в НКРЯ) длиной от одного до 3-4-х предложений[9], соединенный посредством гиперссылки с клипом – эпизодом фильма, содержащим соответствующий звучащий текст; эту единицу выдачи в дальнейшем мы называем *клипотекстом*;

Заметим, что текст, соответствующий клипу, будет даваться в стандартной русской орфографии, без использования какой-либо системы транскрибирования, что связано с а) базовыми характеристиками НКРЯ, в рамках которого осуществляется данный проект, – в НКРЯ все тексты выдаются в стандартной русской орфографии, что делает их доступными абсолютно всем пользователям; с б) отсутствием необходимости в транскрибировании звучащего текста, поскольку пользователю доступен сам первоисточник, звучащий текст – соответственно, любой пользователь может при необходимости затранскрибировать его сам, используя свою собственную транскрипцию и с той степенью подробности, которая удовлетворит его самого. Тем самым предложенная система выдачи звучащего текста снимает с создателей КЗРР одну из самых обременительных проблем – проблему разработки последовательной системы транскрибирования звучащего текста.

2) *клип* – эпизод фильма, не содержащий звучащей речи, но содержащий определенный жестовый материал.

Выдача примеров в КЗРР возможна по двум типам запросов.

Метатекстовый запрос. Пользователь может запросить клипы и клипотексты, которые обладают теми или иными метатекстовыми (т. е. не связанными с конкретным словом, морфемой, граммемой, значением) характеристиками (о предварительном наборе таких характеристик – ниже). Дальше стратегия работы пользователя с полученными при выдаче единицами строится в зависимости от его потребностей – он может работать с клипами и клипотекстами как со зрительными и звуковыми объектами, или может на базе полученного материала сформировать пользовательский подкорпус, чтобы на его основе осуществлять запросы второго типа, текстовые.

Текстовый запрос. Строится как стандартный для НКРЯ запрос по стандартным характеристикам: запрос от точной формы, от лексемы, от морфологической характеристики, от семантической характеристики, а также от комбинации этих параметров. В качестве ответа на такого рода запросы выступает набор клипотекстов, с которыми пользователь имеет возможность вести работу, – аналогично тому, как сейчас ведется работа с контекстами, полученными по стандартным запросам в НКРЯ.

Обратим внимание на то, что по текстовым запросам пользователь получает только клипотексты, а по метатекстовым – и клипотексты, и клипы.

3.1.1. Типы текстовых запросов

Типы текстовых запросов в КЗРР не отличаются от тех, которые разработаны для устного подкорпуса НКРЯ, т. е. это стандартные запросы (лексические, морфологические и семантические), принятые в НКРЯ в целом, в сочетании со специфическими возможностями, предусмотренными для устного подкорпуса. Среди последних следует прежде всего отметить возможность поиска по социологическим характеристикам. К последним относятся 1) гендерные характеристики[10] и 2) характеристики по году рождения говорящего. В устном подкорпусе уже сейчас каждому слову, наряду с морфологическими и семантическими, приписываются гендерные и возрастные характеристики (в случае, если они известны), соответственно, возможны запросы типа «использование определенной лексемы в речи мужчин», «использование определенной словообразовательной модели в речи женщин того или иного года рождения» и под. Кроме того, социологическая разметка в подкорпусе «Речь кино» позволяет строить лексические, морфологические и семантические запросы к речи определенного актера, к речи группы актеров, год рождения которых попадает в определенный период, и нек. др.

Все эти типы запросов, как предполагается, будут сохранены в КЗРР, просто, в отличие от устного подкорпуса НКРЯ, при этом полученные на выдаче контексты будут клипотекстами.

---

[9] Следует отметить, что чрезвычайно чувствительной является проблема разбиения фильма на отдельные блоки. На данном этапе мы принимаем решение считать отдельной единицей описания минимальный относительно законченный блок текста/видеоряда, при вычленении которого не приходится насильственно прерывать речь персонажей и параллельный жестовый ряд. Мы надеемся, что будущее устройство КЗРР позволит пользователю, в случае, если ему потребуется расширить контекст выдачи, обратиться к предыдущему и последующему блоку/эпизоду.

[10] Поскольку на начальном этапе работы материалом для КЗРР, как сказано выше, послужат кинотранскрипты, то в качестве говорящих выступают актеры, а гендерные характеристики принимают несколько своеобразный вид, а именно: помимо очевидных характеристик «мужской» и «женский», используются характеристики «мужской-женский» (актер-мужчина, играющий женскую роль) и, соответственно, «женский-мужской» (актриса, играющая мужскую роль).

019-Grishina_E:_Layout 1 18.05.2008 14:36 Page 128

3.1.2. Типы метатекстовых запросов

Каждый клип и клипотекст в составе КЗРР обладает рядом метатекстовых характеристики.

1) ***Метатекстовые характеристики всего фильма целиком*** (и, соответственно, каждого составляющего этот фильм клипа/клипотекста): авторы (режиссер, автор(ы) сценария, автор исходного текста – при экранизации), год рождения авторов, название, год создания фильма, место расположения киностудии, жанр (все эти характеристики и сейчас используются при описании кинотранскриптов в устном подкорпусе НКРЯ).

2) ***Метатекстовые характеристики клипа, имеющего звуковую, но не имеющего жестовой составляющей*** (чаще всего это касается клипов, содержащих т. н. «голос за кадром», но возможны и случаи, когда персонажи переговариваются в темноте, или в отдалении, так что жесты попросту недоступны для восприятия, в отличие от речи; впрочем, встречается некоторое количество эпизодов, в которых при полноценности речевой составляющей жесты максимально редуцированы).

В этом случае, как нам представляется, клип должен характеризоваться по следующим параметрам (очевидно, что если в клипе есть несколько фраз, то каждая из них получает свой набор характеристик, которые плюсуются при характеристике целого клипа):

– **тип речевых действий,** имеющихся в клипе (предварительно выделяются следующие речевые действия: *аргумент, баюканье, благодарность, брань, возражение, вопрос общий, вопрос частный, вопрос косвенный*[11]*, восклицание, жалоба, запрет, заявление, знакомство, зов, извинение, инструкция, клятва, команда, комментарий, комплимент, констатация, молитва, незнание, обвинение, обещание, обращение, объявление, объяснение, оскорбление, отказ, отрицание, пароль, перечисление, пересказ, подсказка, поздравление, порицание, поучение, похвала, похвальба, предложение, предостережение, предсказание, предупреждение, приветствие, приглашение, признание, призыв, приказ, проводы, проклятье, просьба, прощание, разрешение, рапорт, раскаяние, распоряжение, рассказ, реклама, сентенция, соболезнование, совет, согласие, сообщение, торг, торопить, тост, требование, уговор* (когда персонажи договариваются о чем-то)*, уговоры* (когда один персонаж уговаривает другого что-л. сделать)*, угроза, указание направления, упрек, успокаивать, утверждение, шутка*)[12]. Очевидно, что этот список достаточно велик, однако вполне возможно, что он редуцируется в процессе конкретной работы над разметкой материала, – некоторые из перечисленных речевых действий останутся обязательно, а некоторые, например, *торг, реклама* могут и не потребоваться в виду того, что описываемый текст может оказаться слишком небольшим для таких укрупненных характеристик. В любом случае, вопрос окончательной доводки данной позиции в метатекстовой характеристике может ставиться не априори, а только в процессе работы с конкретным материалом;

– **полнота осуществления речевого действия** (предварительно: действие может оцениваться как *полное*, как *незаконченное* – когда речевое действие добровольно прекращается говорящим, как *прерванное* – когда речевое действие прерывается внешним фактором, включая *автопрерывание*, как *продолженное* – когда слушающий продолжает реплику говорящего; кроме того, здесь же отмечаются случаи *наложения* реплик, а также *вопросы, оставшиеся без ответа*);

– **манера говорения** (здесь могут быть выделены *нормальная речь, крик, шепот, пение, речь с дефектами дикции, ненормально быстрая речь, диктовка, скандирование, декламация, голос за кадром, дубляж*);

– **наличие повторов** (здесь могут быть выделены *однократный–многократный* повтор, *однословный–неоднословный, повтор с интенсификатором, повтор с разной интонацией, переспрос, передразнивание, цитирование,* а также *эхо* – повтор со сменой говорящего и с сохранением типа речевого действия)[13];

– **наличие междометий и вокальных жестов**[14]

(имеются в виду не лексикализованные междометия типа *ах, ох, эх, ай, ой* и под. – включающие их клипотексты могут быть получены посредством обыкновенного лексического запроса через текстовый вход в КЗРР, – а междометия, которые не получили стандартного орфографического воплощения, т.е. разного рода эканья, меканья (заполнители пауз, по терминологии А. Д. Шмелева, см. Шмелев 2005), или маркеры хезитации (Подлесская, Хуршудян 2006), а также причмокивания, цоканье языком, физиологически немотивированные сплевывания (то, что в письменных текстах фиксируется как *тьфу*), свисты (недоуменный, имитирующий

[11] Ср. (Кустова 2007).

[12] При формировании списка использованы работы (Вежбицкая 2007), (Гольдин 2007), (Китайгородская, Розанова 2005), (Шмелева 2007), а также результаты предварительного анализа киноматериала.

[13] В принципе, в настоящий момент в НКРЯ в целом и в устном подкорпусе в том числе, благодаря работе программистов Н. Григорьева и А. Аброскина, возможен поиск редупликаций самого разного свойства, однако по очевидным причинам мы не можем запросить в НКРЯ «редупликацию вообще», без привязки к конкретной лексеме, граммеме и т. д. Что касается КЗРР, то единицей в нем является достаточно компактный объект, и информация о наличии/отсутствии в нем повторов и их типе может быть вынесена в метаописание целого клипотекста.

[14] О последних см. (Шаронов 2006).

019-Grishina_E:_Layout 1 18.05.2008 14:36 Page 129

быстрое движение (в письменных текстах иногда фиксируется как *фьють, фьюить*), подзывающий и др.) и т.д.);

– **количество говорящих** в клипе;

– **пол говорящих** (*мужской, женский, смешанный*);

– **язык** (*русский, с акцентом, иностранный, квазиязык, тайный язык*);

3) ***Метатекстовые характеристики клипа, имеющие жестовую, но не имеющие звуковой (речевой) составляющей*** (это касается очень частых в кинематографе эпизодов, отражающих неречевое поведение персонажей, а также случаи, когда речь, фактически сопровождая действия героев, реально зрителю недоступна, например, просто выключена).

В данном случае речь не идет о характеристике клипа как целого – напротив, метатекстовыми атрибутами в соответствии с перечисленными ниже параметрами снабжается каждый жест, который разметчик выделяет в клипе. Соответственно, по жестикуляционным параметрам каждый клип получает несколько метатекстовых описаний.

На данный момент логичными нам представляются следующие параметры описания жестов в КЗРР (предложенный ниже набор параметров полностью основан на работах Крейдлин 2004 и Григорьева и др. 2001 – идеи и разработки, изложенные в этих книгах, были лишь адаптированы нами к реальным обстоятельствам, обычно сопровождающим работу над созданием мало-мальски объемных корпусов):

– **социологические параметры** (имя, пол, возраст актера (возраст не точный, а приблизительный – *ребенок, подросток, молодой человек, взрослый, старый*[15]), пол и возраст персонажа);

– **орган, осуществляющий жест** (рука, голова, туловище, нога);

– **активный орган** (рука, голова, кисть, подбородок, глаза и т.д.);

– **пассивный орган** (орган тела говорящего, являющийся необходимой составной частью жеста, но не являющийся активным/движущимся органом, например, *грудь* при жесте «сложить руки на груди»);

– **адаптор** (необходимая составная часть жеста, не являющаяся частью тела жестикулирующего, т.е. своего рода отчужденный от тела жестикулирующего пассивный орган – например, *одежда* при жесте «поправить пиджак», *собеседник* (*голова*) при жесте «погладить по голове», *внешний объект* при жесте «показать пальцем» и т.д.);

– **направление движения** активного органа (обычно это касается таких органов, как голова и руки, – здесь направления задается с помощью наречий *вверх, вперед* (=вдоль направления взгляда жестикулирующего), *назад, вбок, вниз, сверху, спереди, сзади, сбоку, снизу, по кругу, горизонтально, вертикально* и их комбинаций);

– **кратность жеста** (однократный–многократный);

– **название жеста** (в случае, если жест не отрефлектирован в языке, используются условные названия посредством типичных речевых формул, сопровождающих этот жест, – типа жеста *«стоп»*, описанного в цитированных выше исследованиях по русским жестам, или жеста *«иди»* – распоряжения, сделанного с помощью однократного движения подбородка вперед, и др.)[16];

– **тип жеста** (здесь предварительно выделяются следующие типы: 1) *жесты внутреннего состояния* – эмоции, ментальные состояния, 2) *дейктические жесты* – включают в себя, по терминологии Г. Е. Крейдлина, дейктические эмблемы, дейктические иллюстраторы, 3) *изобразительные жесты* – с помощью которых жестикулирующий показывает форму, количество, направление, расстояние, иллюстрирует значение некоторых слов, 4) *регулирующие жесты* – регулируют ход общения и поведение участников коммуникации, 5) *этикетные* – приветствия, прощания, извинения и под., 6) *декоративные* – прихорашивания, оправление одежды, 7) *условные*, или *символические* – жест «на ять», жест ОК, 8) *корпоративные* – пионерский салют, молитвенные позы, бандитская распальцовка, 9) *жесты – речевые действия* – клятва, согласие, несогласие, 10) *физиологические жесты* – зевать, почесываться, 11) *поисковые* – искать кошелек в карманах, вспоминать слово, выбирать выражение, 12) *риторические* – подчеркивающие и иллюстрирующие ритм и отдельные компоненты содержания речи, 13) *пейоративные жесты*);

– **тип коммуникативного действия** (этот параметр существен а) для *этикетных жестов* – именно здесь указывается, в какой именно ситуации осуществляется данный этикетный жест (приветствие, извинение, знакомство, прощание и под.), б) для *жестов – речевых действий* – именно здесь указывается, какое именно речевое действие совершается данным жестом (согласие, отрицание, угроза, утешение, клятва и под.), в) для *пейора-*

[15] Напомним, что точный возраст актера может быть получен из базы данных корпуса на той же странице, где расположена поисковая форма для текстовых запросов

[16] Отметим, что в названии жеста, которое содержит глагол, должна соответствующим образом отражаться кратность жеста – для однократных жестов используется совершенный, для многократных – несовершенный вид (таким образом, например, различаются кивнуть и кивать)..

019-Grishina_E:_Layout 1 18.05.2008 14:36 Page 130

тивных жестов – указывается, какой именно тип бранного жеста (‘дурак’, ‘сумасшедший’ и под.) используется);
– **тип внутреннего состояния** (очевидным образом параметр имеет отношение только к *жестам внутреннего состояния* и описывает эмоциональные и ментальные состояния жестикулирующего – нежность, удивление, радость, задуматься, догадаться и под.);
– **наличие удлинителя** (наличие необязательного предмета, посредством которого осуществляется жест в данной ситуации, например, удлинитель «головной убор» при жесте «прижать руку к груди», если жестикулирующий держит шляпу в той руке, посредством которой осуществляется жест);
– **наличие спойлера,** или **редуктора** (наличие предмета, который мешает исполнить жест в классическом варианте, например, указание ногой в случае, если руки и голова заняты);
– указание на то, сопровождается жест **улыбкой, смехом, плачем** или нет;
– **полнота жеста** (*полный* жест имеет полный цикл осуществления, *прерванный жест,* в том числе и *автопрерывание*, – прерывается внешними обстоятельствами, *трансформированный жест* по ходу своего осуществления превращается в другой жест);
– **аутентичность жеста** (является ли жест *притворным*; является ли жест *пародией*, *передразниванием* или *зеркальным повторением* чужого жеста).

4) ***Метатекстовые характеристики клипов, имеющих как речевую, так и жестовую составляющую,*** естественным образом являются суммой пунктов 2) и 3).

4. Типы задач, которые могут решаться с помощью КЗРР
4.1. Задачи класса ‘текст звук‘

Естественно, мы не беремся перечислить в данной статье все задачи, которые можно будет решать с помощью КЗРР, – если бы такое перечисление было возможно, то за создание этого корпуса, как представляется, не стоило бы и браться. Хороший корпус, по-видимому, в значительной степени должен жить своей жизнью, которую невозможно было предугадать на стадии его проектирования. Однако некоторые типы задач обозначить все-таки можно.

Прежде всего, КЗРР позволит ставить акцентологические, фонетические и орфоэпические задачи.

1) Как известно, в настоящее время в рамках НКРЯ создается акцентологический подкорпус (или, официально, корпус «История русского ударения»). Этот подкорпус включает в себя русскую силлабо-тоническую поэзию, в которой размечены сильные доли, а также кинотранскрипты, в которых словесное ударение расставлено в соответствии с *реальным* произношением. С внутрисловными ударениями в кинотранскриптах особенных проблем нет, они расставляются достаточно однозначно. Не вызывают трудностей и случаи переноса ударения с полнозначного слова на проклитику. Однако в ряде случаев ситуация с ударениями достаточно неоднозначна – прежде всего это касается сочетания служебных частей речи и местоимений, которые в разных позициях, сочетаниях и смысловых вариантах могут быть то ударными, то безударными. Например, фразы типа *Где он? – Вот он! Что это?* могут произноситься как с отчетливо безударными *он, это* (вплоть до полной редукции гласных, в том числе и потенциально ударных), так и со слабым ударением. Естественно, человек, профессионально занимающийся ударением клитик в современном русском языке, будет испытывать настоятельную необходимость в такого рода точках обращаться к реальному звучанию.

2) Очевидным образом звучащий корпус будет неплохим подспорьем для исследователей и преподавателей фонетики и орфоэпии. Более того, поскольку корпус, как предполагается, будет достаточно сбалансирован с точки зрения хронологии и, кроме того, для значительного числа исполнителей будут указаны года рождения, то на корпусе можно будет ставить своего рода исторические задачи – рассматривать те или иные фонетические явления в их истории от 1930-х годов до сегодняшнего дня. Уже самое первое приближение к использованию звуковой составляющей кинематографического корпуса показывает, что, например, в использовании частицы *вот* в фонетическом варианте [от], т. е. без начального согласного, есть некоторые хронологические закономерности – в фильмах, снятых до 1961-го года, такой произносительный вариант встречается существенно чаще, чем в фильмах, снятых позже[17]. Безусловно, интересно было бы проследить, к примеру, историю произнесения сочетаний заднеязычных (*буХГалтер, К Кому* и под.), стяжений типа *када, тада* (‘когда’, ‘тогда’). И так далее. Фонетисты и специалисты по орфоэпии могут сделать этот ряд примеров практически бесконечным.

3) Для ряда конструкций возможно исследование эмфазы самого широкого свойства (рематической, смысловой и т. д.). В частности, например, интересно распределение логического, фразового и рематического ударения в такой сугубо разговорной конструкции, как *Что, Р, что ли?*, где *Р* может быть равно как одному слову, так и целой фразе.

[17] См. об этом (Гришина 2008).

019-Grishina_E:_Layout 1 18.05.2008 14:36 Page 131

4.2. Задачи ‘текст жест‘

В рамках КЗРР можно будет ставить, среди многих прочих, вопрос о связи тех или иных жестов с определенным лексемами, граммемами и семантическими множителями, например, исследовать риторические жесты, подчеркивающие указания, в их отношении к указательным и неуказательным единицам речи, использованным в соответствующей фразе/комплексе фраз. Можно исследовать степень обязательности связи того или иного жеста с определенной синтаксической конструкцией. Например, при употреблении конструкции *Вот + вопросительное местоимение* (*Вот как…, Вот где…, Вот почему… Вот о чем…* и т.д.) говорящий часто производит характерное движение подбородком вперед с одновременным поднятием бровей и иногда однократным закрытием глаз. Возникает вопрос, обязательно ли это мимическое движение и почему именно оно используется вместе с данной конструкцией? Представляется, что корпус окажется полезным для решения такого рода проблем.

4.3. Задачи ‘звук‘, ‘жест‘, ‘звук жест‘

Возможности постановки тех или иных задач, связанных с не с текстовыми запросами, а лишь со звуковыми или жестовыми, полностью определяются той разметкой, которая предусматривается для клипов и клипотекстов. Так, на систематическую основу можно будет поставить изучение русской интонации, связанной с разными типами речевых действий, изучение типов повторов, прерываний речи, поведение и семантику нелексикализованных междометий и проч. Что касается жестов, то можно было бы обозначить множество русских этикетных, риторических и т.д. жестов, способы и вариации их осуществления, и т. д. Не будучи специалистами в области жестовой коммуникации, мы не беремся продолжить этот ряд, но подозреваем, однако, что он достаточно длинен. Кроме того, возможно соотнести между собой в запросе разметку клипа как звукового файла и разметку этого же клипа как совокупности жестов. Тогда можно было бы, например, ставить проблему соотношения тех или иных типов речевых действий с теми или иными типами жестов.

Разметка внутреннего состояния говорящего и/или жестикулирующего делает корпус незаменимым инструментом в психолингвистических исследованиях, в частности в изучении способов жестового и речевого выражения эмоциональных и ментальных состояний (именно в сфере описания эмоций мультимедийные корпуса широко используются в зарубежной корпусной лингвистике).

Отдельного анализа требуют возможности использования КЗРР в процессе обучения русскому языку. Очевидно, что такие возможности есть, и они достаточно разнообразны, но этот аспект использования корпуса требует отдельного специализированного обсуждения.

***

На этом можно завершить предварительное описание проекта КЗРР (Корпус звучащей русской речи). Представляется, что нам удалось сформулировать этот проект достаточно законченно и целостно, чтобы, по крайней мере, начать его обсуждение. Добавим, что на настоящий момент нами уже практически разработана технологическая цепочка подготовки материалов для КЗРР – здесь не место ее излагать, однако уже сейчас ясно, что этот проект вполне реален и при благоприятных экстралингвистических обстоятельствах может быть воплощен в жизнь в ближайшие годы.

Список литературы

1. Вежбицкая 2007 – Вежбицка Анна. Речевые жанры (в свете теории элементарных смысловых единиц) // Антология речевых жанров: Повседневная коммуникация. М.: Лабиринт, 2007. С. 68-80.

2. Волошинов 1930 – Волошинов В.Н. Конструкция высказывания // Литературная учеба, 1930, № 3. С. 65-87.

3. Гольдин 2007 – Гольдин В.Е. Имена речевых событий, поступков и жанры русской речи // Антология речевых жанров: Повседневная коммуникация. М.: Лабиринт, 2007. С. 90-102.

4. Григорьева и др. 2001 – Григорьева С.А., Григорьев Н.В., Крейдлин Г.Е. Словарь языка русских жестов. М–Вена: 2001

5. Гришина 2005 – Гришина Е.А. Устная речь в Национальном корпусе русского языка // Национальный корпус русского языка: 2003–2005. М.: 2005

6. Гришина 2007 – Гришина Е.А. О маркерах разговорной речи (предварительное исследование подкорпуса кино в Национальном корпусе русского языка) // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2007» (Бекасово, 30 мая – 3 июня 2007 г.). С. 147-156

7. Гришина 2008 – Гришина Е.А. Частица *вот*: варианты, используемые в непринужденной речи // Что проис-

019-Grishina_E:_Layout 1 18.05.2008 14:36 Page 132

ходит в современном русском языке (в свете данных языковых корпусов). Slavica Helsingiensia (в печати)

8. Земская и др. 1981 – Земская Е.А., Китайгородская М.В., Ширяев Е.Н.. Русская разговорная речь. Общие вопросы. Словообразование. Синтаксис. М.: Наука, 1981.

9. Капанадзе 1986 – Капанадзе Л.А. Разговорная речь и киноязык // Л.А. Капанадзе. Голоса и смыслы. Избранные работы по русскому языку. М.: 2005. С. 228-231.

10. Кибрик, Подлесская 2003 – Кибрик, А.А., Подлесская, В.И. К созданию корпусов устной русской речи. // НТИ. Сер. 2. 2003, № 10. С. 5–12.

11. Кибрик и др. 2007 – Кибрик А. Е., Архипов А. В., Даниэль М. А., Кодзасов С. В., Майерс Т., Нахимовский А.Д. Технологии обработки языковых данных в документировании малых языков // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2007» (Бекасово, 30 мая – 3 июня 2007 г.). С. 231-235.

12. Китайгородская, Розанова 2005 – Китайгородская М.В., Розанова Н.Н. Речь москвичей: Коммуникативно-культурологический аспект. М.: Научный мир, 2005.

13. Кляус 2004 – Кляус В.Л. К методике видеофиксации фольклора. http://folk.pomorsu.ru/publicat/_complsobir/vlklaus.html

14. Кустова 2007 – Кустова Г.И. Косвенный речевой акт вопроса как средство речевой агрессии и негативной оценки в русской разговорной речи // Культура русской речи. I Международная научная конференция. 15–17 октября 2007 года.

15. Крейдлин 2004 – Крейдлин Г.Е. Невербальная семиотика. М.: 2004.

16. Мороз 2003 - Мороз А.Б. Из опыта работы над базой данных «Традиционная культура Русского Севера (Каргополье)»// Актуальные проблемы полевой фольклористики. Вып. 2. М . 2003. С. 85-99.

17. Подлесская, Кибрик 2004 – Подлесская, В.И., Кибрик, А.А. Транскрипция устного дискурса для нужд корпусных исследований. // Труды международного семинара «Диалог 2004» по компьютерной лингвистике и ее приложениям. Верхневолжский, 2–7 июня 2004 г. http://www.dialog-21.ru/Archive/2004/Podlesskaja.htm

18. Подлесская В.И., Хуршудян В.Г. О лексических маркерах хезитации в спонтанной речи: уроки армянского. // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.). С. 429-439.

19. Русская разговорная речь // М.: 1973 (РРР 1973).

20. Савчук 2007 – Российские разработки корпусов устной речи (Russische Phonokorpora-Modelle) // 1. Symposium „Die phonetisch-phonologischen, orthoepischen und orthographischen Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen“ Graz, 12.–14. April 2007 (в печати)

21. Савчук 2008 – Савчук С.О. Устный корпус: состав и структура. // Национальный корпус русского языка: 2006–2008 (в печати)

22. Седов 2007 – Седов К.Ф. Человек в жанровом пространстве повседневной коммуникации // Антология речевых жанров: Повседневная коммуникация. М.: Лабиринт, 2007. С. 7-38.

23. Шаронов 2006 – Шаронов И.А. Эмоциональные междометия и вокальные жесты // Русский язык сегодня. Вып. 4. Проблемы языковой нормы. 2006. С. 605-617.

24. Шмелев 2005 – Шмелев А.Д. «Заполнители пауз» как коммуникативные маркеры. // Язык. Личность. Текст. М.: 2005.

25. Шмелева 2007 – Шмелева Т.В. Модель речевого жанра // Человек в жанровом пространстве повседневной коммуникации // Антология речевых жанров: Повседневная коммуникация. М.: Лабиринт, 2007. С. 81-89.

26. Якубинский 1923 – Якубинский Л.П. О диалогической речи // Русская речь. Пг., 1923. С. 96-194.

27. Braun et al. 2007 – Sabine Braun, Ylva Berglund Prytz, Kurt Kohn and Pascual Pérez-Paredes. Multimedia Corpora for Applied Linguistic Contexts // Corpus Linguistics 2007. Book of Abstracts. Birmingham, 2007. P. 22.

28. Clavel et al. 2006 – Clavel Ch., et al. Fear-type emotions of the SAFE Corpus: annotation issues // 5th International Conference on Language Resources and Evaluation. 22-28 May 2006. Genoa, Italy. Conference Abstracts. P.76.

29. Devillers et al. 2006 – Devillers L., et al. Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches // 5th International Conference on Language Resources and Evaluation. 22-28 May 2006. 22-28 May 2006. Genoa, Italy. Conference Abstracts. P.76.

30. Grishina 2006 – Grishina Elena. Spoken Russian in the Russian National Corpus (RNC) // LREC2006 5th International Conference on Language Resources and Evaluation. 22-28 May 2006. Genoa Italy. Proceedings. P. 121-124.

31. Grishina 2007 – Grishina E. Text Navigators in Spoken Russian. / Proceedings of the workshop “Representation of Semantic Structure of Spoken Speech” (CAEPIA’2007, Spain, 2007, 12-16.11.07, Salamanca), Salamanca, 2007. P. 39-50.

020-Desjatova_A:_Layout 1 18.05.2008 14:37 Page 133

КОНСТРУКЦИЯ С ТВОРИТЕЛЬНЫМ ФОРМЫ «X Y-OM»
DESCRIBING SHAPE: INSTRUMENTAL CONSTRUCTION «X Y-OM»

Десятова А.В. *(patine@gmail.com), Российский государственный гуманитарный университет*
Ляшевская О.Н. *(olesar@mail.ru), Институт русского языка им. В.В. Виноградова РАН*
Махова А.А. *(discourse@yandex.ru), МГУ им. М.В. Ломоносова*

В работе проводится исследование семантики русской конструкции с творительным формы типа *хвост кольцом, сложить губки бантиком*. Пространственная семантика конструкции описывается в терминах топологических классов (Talmy 2000, Рахилина 2000): анализируются явления совпадения топологических классов и их взаимной аккомодации.

1. Введение

В рамках изучения топологических (геометрических) характеристик предметных имен[1] нами была поставлена задача рассмотреть, каким образом форма объектов описывается в русском языке с помощью конструкции с творительным падежом (типа *хвост кольцом*).

Прежде чем перейти к описанию материала, необходимо отметить, что в некоторых случаях примеры трудно однозначно охарактеризовать с грамматической точки зрения. Возникает два вопроса. Во-первых, сложно провести границу между существительными в творительном падеже и наречиями, мотивированными существительными в творительном падеже (типа *торчком*). Неслучайно В.В. Виноградов считал, что формы творительного падежа образа действия (а также формы творительного сравнения и времени) могли бы обособиться в самостоятельный падеж, так как в современном русском языке они являются «гибридными наречно-субстантивными образованиями» (1972: 143). Соответственно, в грамматиках примеры конструкций с творительным формы попадают разные разделы. Компонент Y может быть грамматически охарактеризован либо как «наречие образа действия с частным значением сравнения, уподобления (*ёжиком*)» [КРГ, 260, РГ сюда же относит словоформу *столбом*], либо как «существительное в форме творительного падежа, обозначающее предмет, придающий действию известный вид, форму, облик, то есть передающий значение образа действия (*уложить косы короной*)» [КРГ, 382].

Во-вторых, эта конструкция упоминается в грамматиках как пример слабого управления в описании глагольных конструкций типа *свернуть хвост кольцом* (см. также пример выше), и в то же время мы встречаем другие ее примеры, описываемые как «примыкание тв. п. к существительным конкретных значений, обычно определяемых со стороны внешних признаков или свойств, качеств: *сапоги раструбами*, *галстук шнурочком*…» [РГ, § 1838.2]. Так же как и первый вопрос, дискуссию о статусе глагола в данной конструкции[2] мы бы хотели оставить в стороне – более того, лексические ограничения на глагол не являются специальным предметом нашего исследования. Таким образом, в данной работе мы понимаем «конструкцию с творительным формы» в расширительном значении: вторым элементом конструкции могут быть имена существительные в творительном падеже или наречия, мотивированные таким творительным падежом; кроме того, конструкция может включать глагол (*X V Y-ом*), или же элементы X и Y могут быть связаны непосредственно (примыкание).

Также надо подчеркнуть, что рассматриваемая нами конструкция - это более узкая и специальная группа, чем конструкция с творительным образа действия, и она практически не описывается в лингвистических работах. Часто указывается, пожалуй, лишь тот факт, что эта конструкция имеет дополнительное значение сравнения. Г.А. Золотова отмечает лексическую ограниченность синтаксемы тв. п. «со значением *предмета*

[1] Работа выполнена в рамках проекта «Топологические типы русских предметных имен: корпусное исследование» при поддержке РГНФ, грант № 07-04-00240а.

[2] В некоторых исследованиях эта конструкция функционально связывается с разговорной речью и опущением глагола. «В разговорном языке можно выделить группу устойчивых субстантивных метафорических конструкций с наречиями, мотивированными существительными в форме творительного падежа: нос картошкой (пуговкой), хвост крючком (кренделем), губы бантиком, борода лопатой, грудь колесом, ноги циркулем. Эти словосочетания образовались в результате утраты первой части именного глагольного сказуемого: нос вырос картошкой, хвост свернулся крючком, борода спускается лопатой и т. д.: «Широкая густая борода, спускавшаяся лопатой, была еще светлее головных волос» (Ф.Достоевский). Функциональное значение метафор, входящих в их состав, менее разнообразно, чем в случае эксплицитно представленных глагольных словосочетаний, и ограничивается характеристикой по форме» [Глазунова 2000].

020-Desjatova_A:_Layout 1 18.05.2008 14:37 Page 134

сравнения, уподобления в качестве предикативной или атрибутивной характеристики некоторых предметных имен» [Золотова 2001, 243].

Какие ограничения накладываются в русском языке на эту конструкцию и с чем это связано? Чтобы ответить на этот вопрос, мы рассмотрим лексико-семантические и топологические аспекты этой конструкции.

2. Материал исследования

Анализ конструкции строится по преимуществу на материале Национального корпуса русского языка[3]. Отправным пунктом исследования послужил список слов, имеющих словарную помету, указывающую на то, что они могут описывать форму предметов. С помощью корпуса был составлен список словосочетаний и примеров, который показал, что ограничения на использование слов в этой конструкции очень существенны, и большинство слов, даже имеющих пометку «форма», не могут быть использованы в этой конструкции в позиции Y (например, *звезда, конус, линия*). В дальнейшем список словосочетаний был расширен за счет примеров из грамматик русского языка и примеров, найденных с помощью поисковой системы Яндекс.

Поскольку наше исследование было нацелено на то, чтобы обнаружить, как инструментальная конструкция проявляет семантику формы индивидуальных предметов, мы исключили из рассмотрения словосочетания с названиями множественных объектов (*апельсины горкой, книги стопкой*) и веществ (*дым столбом, тесто колбасками*); а также обособленную группу ‘еда + форма нарезки’ (*картофель брусочками, мясо кубиками*).

3. Ограничения конструкции: лексико-семантические группы в позиции X и Y

Итак, слова каких семантических групп представлены в инструментальной конструкции формы? В позиции X наиболее широко представлена семантическая группа «части тела», например, *руки* (*крестом* – 28 примеров[4]), *грудь* (*колесом* - 18), *ноги, голова, живот,* причем встречаются не только части тела человека, но и части тел животных: *хвост* (*трубой* – 16), *хобот, клюв, загривок, крылья.* Внутри этой группы наиболее частотны части лица человека, и в первую очередь *губы* (*трубочкой* – 49 примеров, *бантиком* – 11, *сердечком* – 8 и др.):

(1) *Лицо у него было доброе,* **губы трубочкой**, *словно он все время что-то насвистывал* [Р. Солнцев Роман. Полураспад. Из жизни А. А. Левушкина-Александрова, а также анекдоты о нем // «Октябрь», 2002];

(2) *Савва обиженно вытянул* **губы трубочкой** [М. Баконина. Школа двойников (2000)].

Другие части лица человека также представлены широко, особенно *борода/бородка* (*клином/клинышком* - 72), *нос, брови.*

Гораздо реже в позиции X встречаются имена одежды (*юбка, платье, сарафан, штаны, сапоги, рубаха, халат*), других артефактов (*чемодан, колода, тетрадь, браслет, шланг*) и – в единичных случаях - природных объектов, а именно *тропа, река, дорога*:

(3) *Я смотрю на огромную оцепенелую округу, в которой живет и пульсирует в холмах лишь* **дорога петлями** - *гигантский кишечник во вскрытой брюшной полости Иудейской пустыни* [Д. Рубина. Камера наезжает (1993-1994)];

(4) **Дорога петлями** *падала с обрыва* [И. Ильф, Е. Петров. Золотой теленок (1931)].

Семантические группы в позиции Y – это, в первую очередь, артефакты, например, *рупор* (*ладони рупором* – 36), *колесо* – (*грудь/ноги/спина колесом* – 30):

(5) *Выглядел он так, как и полагается оперному басу, - рост под два метра и* **грудь колесом** [А. Ткачева. Приворот (1996)];

(6) *По проходу между нарами медленно идет в окружении целой свиты начальник пересылки - легендарный Курило, с* **ногами колесом**, *как у заправского кавалериста, и со стеком в руке* [О. Волков. Из воспоминаний старого тенишевца (1988)].

(7) *Голову прижать к груди и сделать кувырок по руке,* **спине «колесом»** *в прямой плоскости, фото 3* [А. Яшкин. Акробатика в каратэ // «Боевое искусство планеты», 2004].

Ср. также *гармошка, бутылка, серп, пуговка, флажок, жгутик, лесенка, колокол, веер, клещи, замок* и другие.

Реже встречаются ‘природные объекты’, в том числе названия ландшафта (*гора, холм, бугор*), а также еды (*груша, картошка, тыква, яйцо, колбаска*) и животных (*уточка, ежик*):

(8) *Эти* **ноги колбасками**, *ушки в оттопырку,* **коса бубликом** - *все это секс-символ украинской политики* [http://censor.net.ua/go/viewTopic—id—162647 – Яндекс];

[3] При цитировании без специальных указаний – примеры из НКРЯ.

[4] При указании на количество примеров приводятся данные Национального корпуса русского языка.

020-Desjatova_A:_Layout 1 18.05.2008 14:37 Page 135

(9) *Невысокий, глаза затаились,* **нос уточкой** *[Архангельский Александр. 1962. Послание к Тимофею (2006)]*

В единичных случаях в позиции Y встречаются обозначения частей тела и образований на теле (*кишка, горб*):

(10) *Одних гнетет и мучает армейская муштра, на них все сидит мешком, гимнастерка пузырится, пряжка ремня на боку, сапоги на три номера больше,* **шинель горбом***, язык заплетается* [В. Некрасов. В окопах Сталинграда (1946)]

Обращает на себя внимание, что многие имена, занимающие позицию Y, входят также в семантический класс диминутивов (о семантике сходства по форме у диминутивов см. [Спиридонова 1999]). Во многих случаях уменьшительные имена выглядят в данной конструкции предпочтительнее своих словообразовательных коррелятов, ср. *нос уточкой / ?уткой*, *губы трубочкой / ?трубкой, хвост флажком / ?флагом*.

4. Топологические классы имен в позиции X и Y

Поскольку речь идет о форме объектов, то описание данной конструкции будет неполным без топологической характеристики ее составляющих. Топологическая характеристика объекта в лингвистическом понимании – это совокупность его значимых геометрических характеристик. Топологический класс предметных имен – это класс слов, имеющих сходную сочетаемость, например, с прилагательными формы и размера [Рахилина 2000], и описывающих объекты, имеющие общие топологические характеристики. В рассматриваемой конструкции и первому, и второму слову может быть приписана топологическая характеристика, то есть слово может быть отнесено к тому или иному топологическому классу[5].

<u>В позиции X</u> мы выделили следующие классы:
– ‘веревки’ (*волосы, косички, усы, хвост, хобот, шланг*),
– ‘выступы’ (*брови, нос, груди, щеки, рот, губы, борода, ушки, живот*),
– ‘кольца/круги‘ (*браслет, повязка*),
– ‘оболочки’ (*рубаха, сарафан, ткань, штаны, сапоги, голенища*),
– ‘пластины’ (*ладонь, бумага, листок, записка, тетрадь, блин*),
– ‘поверхности’ (*спина, живот*); узким подклассом поверхностей также является группа имен типа *свет, солнце, тень* в переносном значении ‘участок земли, на который (не) падает свет’,
– ‘стержни’ (*руки, пальцы, сигарета*),
– ‘дуги (непрямые стержни)‘ (*брови, крылья*),
– ‘столбы (вертикально ориентированные стержни‘ (*ноги*),
– ‘полосы’ (*дорога, линия, шов*),
– ‘трехмерные объекты неопределенной формы’ (*подушка, диван, книжка, крышка, колода, чемодан*),
– ‘верхние части‘ (*стрижка, прическа, шляпа, волосы*),
– ‘шары’ (*голова*).

Принципиально важно, что одно и то же имя может относиться к нескольким топологическим классам. Это связано с тем, что топологическая характеристика объекта может меняться, если кардинально меняется перспектива его восприятия. Так, *волосы* относятся к топологическому классу ‘веревки’ (*волосы локонами*), но в значении ‘стрижка, прическа’ они соотносятся с особым трехмерным объектом, который может обладать любой формой, и в этом значении мы помещаем имя *волосы* в топологический класс ‘верхние части’ (*волосы шапкой*). Другой очевидный пример: *руки* воспринимаются как ‘стержни‘ в значении ‘arm’ (*руки горбом*) и как ‘пластины‘ в значении ‘hand’ (*руки чашей*); ср. также *грудь* как ‘выступ‘ и как ‘поверхность‘. О других переходах из одного класса в другой будет сказано ниже в разделе 5.

<u>В позиции Y</u> выделяются следующие топологические классы:
– ‘веревки’ (*змея, жгутик, плеть*),
– ‘верхние части’ (*шапка*),
– ‘выступы’ (*бугор, горбинка, горб, пузырь, гора*),
– ‘имена формы’ (*вихор, завиток, загогулина, клин, крючок, закорючка, петля, валик, треугольник, угол*),
– ‘кольца/круги‘ (*кольцо, корона, колесо*),
– ‘вместилища‘ (*бутылки, чаша*),
– ‘пластины и поверхности’ (*стена*, *водопад*),
– ‘полосы’ (*нитка*, *лента, складка*),
– ‘непрямые стержни (дуги)‘ (*полумесяц, дуга*),

[5] Топологическая классификация разрабатывается в настоящее время как часть семантической разметки Национального корпуса русского языка.

020-Desjatova_A:_Layout 1 18.05.2008 14:37 Page 136

– ‘столбы’ (*столб, колбаска*),
– ‘трубы’ (*труба, трубка*),
– ‘шары’ (*шарик, яйцо, ком*),
– ‘объекты характерной формы’ (*груша, корзиночка, флажок, ежик, картошка, тыква, уточка, пуговка, лопата, мочалка, гармошка, колокол, домик*).

Класс ‘объектов характерной формы‘ – особый. Он выделяется нами для имен двухмерных и трехмерных объектов, которые обозначают эталон формы [Кобозева 2000], прежде всего, в данной и других смежных конструкциях – в конструкции с родительным формы (*дуги бровей*) и в конструкции с предлогом *в*, управляющим винительным падежом (*сапоги в гармошку*). По сути, этот класс имеет сдвоенный статус как таксономического (родо-видового, онтологического), так и топологического. В него входят как имена чистой формы (то есть те, у которых абстрактное значение абстрактной формы является главенствующим: *крест, дуга, уголок*), так и имена материальных объектов, которые приобретают значение формы только в указанных конструкциях; для них значение формы – контекстное, идущее от семантики конструкции. Словари указывают рассматриваемое раздвоение значения лишь в некоторых, наиболее очевидных случаях. В частности, слово *сердечко* в рассматриваемой конструкции никогда не встречается в значении ‘часть тела’, но приобретает значение формы (ср. *пироги сердечком*).

Благодаря включению в классификацию ‘объектов характерной формы’, одно и то же предметное имя в позиции X и в позиции Y может относиться к разным топологическим классам: например, слово *жгут* может входит в класс ‘веревки‘ и в класс ‘объекты характерной формы‘; в позиции Y *жгут* реализует вторую характеристику:

(11) *…Закручиваю* **волосы жгутом**, *сворачиваю в узел на затылке или низко у шеи, по настроению, и прокалываю палочкой насквозь, как булавкой* [http://forum.cofe.ru/showthread.php?s=&threadid=2282&perpage=25&pagenumber=6 – Яндекс].

5. Взаимодействие топологических классов в конструкции

Топологические классы слов, встречающихся в позиции X и Y, соотносятся между собой по-разному, в рассматриваемой конструкции мы насчитали около 40 разных комбинаций. Интерес представляет совпадение классов X и Y, например, в (12) и слово *нос*, и слово *горб* относятся к топологическому типу ‘выступы’:

(12) *Глаза блестять як у цыгана,* **нос горбом**, *зубы, як мел, вусы, як у Буденнага. А он ей и отвечаить: «Спасибо тибе, барышня, буду служить тибе верую и правдую!* [Л. Гурченко. Аплодисменты (1994-2003)]

Также можно найти сочетания слов, в которых оба элемента относятся к топологическим классам ‘полосы’ (*дорога лентой*), ‘дуги‘ (*брови дугами/полумесяцем*), ‘столбы’ (*ноги столбами*), ‘шары’ (*голова шариком*).

Если же топологические классы в позиции X и Y не совпадают, то происходят того или иного рода трансформации. В чистом виде потеря топологических характеристик X и приобретение характеристик Y носит единичный характер, в основном под влиянием сильного (в т. ч. глагольного) контекста, ср.:

(13) *Нет-нет да попадаются шланги со слабым каркасом - пока тормозишь умеренно, ничего не заметишь, а нажмешь сильнее -* **шланг** *раздувается* **грушей**! [А. Вайсман. О смелых и умелых // «За рулем», 2004]

– из (двумерного) объекта типа ‘веревки‘ мы получаем трехмерный ‘объект характерной формы‘.

В основном же форма описываемого объекта наследует пространственные характеристики X и Y по определенным законам аккомодации топологических типов. В качестве примера рассмотрим сочетание классов ‘выступ‘ + ‘трехмерный объект (характерной формы)‘, ср. *нос картошкой/грушей/сливой/уточкой* и т.д. Выступ – это частный случай трехмерного объекта, прикрепленного к некоторой поверхности. В результате топологической трансформации выступ остается выступом, приобретая характерные черты Y, вместе с тем, теряются такие топологические характеристики Y, как независимое (или привычное) положение в пространстве.

Вообще, выступ – очень устойчивый топологический тип. Ср., например, *губы*: назовем ли мы их *трубочкой*, *бантиком* или *сердечком* – во всех случаях они останутся выступом, лишь чуть большим по сравнению с некой средней мерой, или нормой. *Борода* может быть описана как *клинышек* или *лопата*, но она не меняет при этом своей топологической характеристики; словосочетание приобретает лишь дополнительное указание на меньший или больший размер выступа и особенность формы: суженная к низу или широкая и длинная, как у Деда Мороза:

(14) *Да и сам обходящий квартиры Дед Мороз в исполнении знакомого мне артиста Фимы Абрамова с каждым годом все больше похож на Клауса: колпак вместо шапки,* **борода** *не широкой* **лопатой**, *а*

020-Desjatova_A:_Layout 1 18.05.2008 14:37 Page 137

струечкой [Р. Арифджанов. Москва католическая // «Столица», 1997.01.06].

Грудь, являясь ‘выступом‘, в сочетании с именами топологического типа ‘кольцо‘ (ср. *грудь колесом*) не меняет существенных особенностей своей формы. Она остается выступом, и при таком типе трансформации, как уже говорилось, опять возникает дополнительное размерное значение: объект выступает больше среднего, ср.:

(15) *А путанка Клавдия Ивановна, пятипудовая женщина, сидела у самовара, распаренная, в тренировочных штанах,* **грудь колесом***, размалеванная и в бигуди* [Л. Измайлов. Объезд по кривой (1988)].

В позиции Y тип ‘выступ‘ также доминирует, ср. сочетание с типом X-а ‘поверхность‘:

(16) *Шерсть дыбом,* **спина горбом***, хвост свечкой - вот-вот кинется* [П.П. Бажов. Сочневы камешки (1937)].

Полной противоположностью ‘выступам‘ являются ‘веревки‘: это крайне неустойчивый топологический тип. Будучи объектом изменяемой формы (который легко гнется, извивается, скручивается, сворачивается в кольцо), веревка старается приобрести топологические признаки, которыми обладает объект, называемый именем Y. В примерах из НКРЯ с существительным *хвост* мы можем встретить *хвост крючком, серпом, калачиком, закорючкой* и мн. др. В таком случае *хвост* ведет себя как представитель топологического класса ‘веревок‘ - то есть как вытянутый объект, не имеющий жесткой формы (с дополнительным топологическим признаком прикрепленности к другому объекту). Но встречается также *хвост трубой, палкой, морковкой, поленом*, ср. также пожелание держать хвост *столбом* или *пистолетом*. Здесь тип ‘веревка‘ взаимодействует с топологическим типом ‘стержень‘ (или его подтипом ‘столб‘), полностью уподобляясь последнему: *хвост* представляется как вытянутый объект жесткой формы (признак гибкости формы теряется, единственное, что сохраняется – это идея прикрепленности к другому объекту). В выражении *хвост кольцом* мы имеем дело с чистой трансформацией: объект-веревка сворачивается и приобретает форму кольца. Наконец, в следующем примере:

(17) *Девочки не укладывают* **косы корзиночкой***, а мальчики не стригутся «под бобрик»* [http://www.mosoblpress.ru/mig/show.shtml?d_id=5515 – Яндекс]

из сочетания типов ‘веревка (прикрепленная)‘ + ‘трехмерный объект характерной формы‘ (*косы корзиночкой*) получается трехмерный объект, прикрепленный к поверхности (т. е. выступ).

Различие в поведении типа ‘выступ‘ и типа ‘веревка‘ вполне объяснимо. Можно предположить, что существует иерархия топологических предпочтений:

изменяемая форма < фиксированная форма,

независимое положение в пространстве < прикрепленность,

стержень < поверхность < трехмерный объект,

и, таким образом, тип ‘выступ‘ удачно совмещает два устойчивых признака прикрепленности и трехмерности. В свою очередь, слабость признака изменяемой формы доказывают также конструкции с именами топологического класса гибких поверхностей и пластин типа *язык трубочкой, газета трубочкой*: здесь плоскость легко трансформируется в жесткий трехмерный объект:

(18) *... две недели тяжелого бронхита, ярко-розовый* **язык трубочкой** *торчит изо рта в натужном захлебывающемся кашле* [А. Боссарт. Повести Зайцева (1998)].

И напротив, сочетание ‘столб‘ + ‘кольцо/круг‘ дает компромиссный тип ‘дуга‘, а не ‘кольцо‘, ср. *ноги колесом.*

Вместе с тем, следует оговорить, что в иерархии «стержень - поверхность - трехмерный объект» не все так однозначно. Дело в том, что существуют стандартные трансформации в обратном направлении: любой физический объект может быть представлен как видимая поверхность, видимый контур и точка. Ср. следующие примеры:

а) ‘выступы’ → ‘поверхности’ (ср. *щеки, живот*: *щеки пузырями, живот горой*, но *лоб/живот складками*, ср. также *губы оборочкой/гармошкой*);

б) ‘выступы’ → видимый контур: ‘кольцо‘ (ср. *уши, ушки*: *ушки торчком*, но *уши петлями*, т. е. уши при виде сбоку имеют контур большой петли):

(19) *Всех их [агентов] снабдили поисковой карточкой: «Рост средний, уши* **петлями***, нос прямой, крупный, глаза выпученные, синие, волосы темные, телосложение нормальное»* [В. Кожевников. Щит и меч (1968)];

в) ‘выступы’ → видимый контур: ‘линия‘ (ср. *нос*: *нос горбинкой* и *нос загогулиной*, например, у боксера; ср. также *губы ниточкой*);

г) ‘поверхность‘ → видимый контур: ‘дуга‘, ср.:

(20) *Сквозь частые прутья нельзя было просунуть лапу, и Василий зловеще шипел, выгибал* **спину дугой** *и*

[6] Заметим, что некоторые имена объектов характерной формы используются исключительно для обозначения видимого контура объектов, ср. брови домиком, ладони шалашиком (линия, повторяющая форму крыши).

020-Desjatova_A:_Layout 1 18.05.2008 14:37 Page 138

вполголоса по-кошачьему ругался [Е. Чеповецкий. Приключения шахматного солдата Пешкина (1986)];[6]

д) ‘трехмерный объект‘ → ‘точка‘, ср.:

(21) *На дне встречаются одинокие голубые звезды, красные, синие, белые <...> Уаскаро остановился, нагнулся, протянул руку, и вот обессиленная* **звезда бусиной** *догорает в его пальцах* [Улья_Нова. Инка (2004)].

6. Дополнительные семантические наращения

Совпадение или аккомодация топологических типов может сопровождаться появлением новых пространственных смыслов, в первую очередь, касающихся размера и ориентации объекта. Так, при совпадении топологических типов в словосочетании *живот горой* (‘выступ‘ + ‘выступ‘) возникает дополнительное размерное значение ‘выступ больше нормы’, которое может быть поддержано контекстом:

(22) *Глаза его, без пенсне, обнаженные, слепые, не мигали, лицо было цвета сухой глины, большой* **живот горой** *ходил под натянутым полотном рубашки* [В. Набоков. Машенька (1926)].

Выше, в разделе 5, мы также приводили примеры подобного эффекта при аккомодации топологических типов (*борода лопатой, грудь колесом*).

Интересная трансформация с предсказанием ориентации объекта происходит в конструкции *ладонь козырьком* (варианты: *рука козырьком, ладошка козырьком*). *Ладонь* из топологического класса ‘пластин‘ переходит в топологический класс ‘выступов‘, прямо уподобляясь козырьку. В наиболее распространенном употреблении данной конструкции ориентация части тела такова, что внутренним ребром она прикасается ко лбу:

(23) *Пристально смотрел с картины Илья Муромец, поставив над глазами* **ладонь козырьком** [Ю. Коваль. Приключения Васи Куролесова (1977)].

Но есть и более редкий тип употребления, когда ориентация одной или обеих ладоней следующая: внутренним ребром ладонь прижата к щеке, а внешним (обычно) к стеклу окна:

(24) *Татьяна вскочила с постели в одной рубашке, подошла к окну и, приложив* **ладони козырьком** *к щекам, стала всматриваться через стекло* [Б. Можаев. Власть тайги (1954)].

Дополнительное значение ориентации присутствует и в выражении *язык трубочкой*: оно обозначает, что положение языка продольное, язык высунут, края завернуты вверх.

Помимо нетривиальных пространственных приращений, конструкция может иметь оценочное значение, причем более распространена отрицательная эстетическая оценка, ср. словосочетания *нос крючком, ноги колесом/циркулем, голова огурцом*, *борода мочалкой*:

(25) «**Нос** был **крючком**, - радостно улыбаясь, сказала соседка, - настоящая ведьма. [Б. Окуджава. Упраздненный театр (1989-1993)];

(26) *От обедов же решительно никогда не отказывался и, принимая хлеб-соль Силы Терентьича, ел истово, ног на стол не клал, божьего дара под стол не кидал, банта всенародно не расстегивал и хозяйскую* **бороду мочалкой** *не обзывал* [М.Е. Салтыков-Щедрин. Сатиры в прозе (1859-1862)].

7. Заключение

Итак, анализ конструкции с творительным формы позволяет выделить лексико-семантические классы слов, встречающихся в этой конструкции, а именно: большинство примеров описывает форму частей тела, существенно реже описывается форма предметов одежды и в единичных случаях – форма предметов. Части тела, как и другие предметы, могут быть охарактеризованы с точки зрения топологии; топологическая классификация обозначающих их имен существительных позволяет системно описать пространственную семантику конструкции. Топологические классы имен по-разному взаимодействуют друг с другом: топологический класс имени X может совпадать с типом имени Y (ср. *живот горою, щеки пузырями*, *тень клином*); целиком перестраиваться в тип Y (ср. *шланг грушей*) или же «аккомодировать», согласовываться, подстраиваться под тип Y, трансформируясь в достаточно предсказуемый вторичный топологический тип (ср. *ноги колесом, спина дугой, хвост трубой*).

Список литературы

1. Виноградов В.В. Русский язык. М. 1972.
2. РГ - Русская грамматика. М. 1980.
3. КРГ - Белоусов В.Н., Ковтунова И.И., Кручинина И.Н. и др. Краткая русская грамматика. М., 1989.

020-Desjatova_A:_Layout 1 18.05.2008 14:37 Page 139

4. Золотова Г.А. Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса. М. Эдиториал УРСС, 2001.

5. Глазунова О.И. Логика метафорических преобразований. СПб., 2000.

6. Кобозева И.М. Как мы описываем пространство, которое видим: форма объектов // А.С. Нариньяни (ред.), Труды международного семинара «Диалог 2000» по компьютерной лингвистике и его приложениям. Т.1. Протвино. С. 155-161.

7. Рахилина Е.В. Когнитивный анализ предметных имен: семантика и сочетаемость. М.: Русские словари, 2000.

8. Спиридонова Н.Ф. Русские диминутивы: проблемы образования и значения // Известия АН, СЛЯ, т. 58, 1999, № 2.

9. Talmy L. How language structures space // Talmy L. Toward a cognitive semantics. V. I. Cambridge, MA: MIT Press, 2000.

021-Dobrovolskiy-Paducheva:_Layout 1 18.05.2008 14:44 Page 140

ДЕЙКСИС В ОТСУТСТВИЕ ГОВОРЯЩЕГО: О СЕМАНТИКЕ НЕМЕЦКИХ ДЕЙКТИЧЕСКИХ ЭЛЕМЕНТОВ *HIN* И *HER*

DEIXIS WITHOUT SPEAKER: TOWARDS THE SEMANTICS OF THE GERMAN DEICTIC ELEMENTS *HIN* AND *HER*

Добровольский Д.О. *(dm-dbrv@yandex.ru), Институт русского языка РАН*
Падучева Е.В. *(elena708@gmail.com), Институт русского языка им. В.В.Виноградова РАН*

Традиционно семантика дейктических элементов *hin* и *her* описывается исходя из позиции говорящего. Это верно, однако, только для канонической коммуникативной ситуации. В неканонической ситуации (когда говорящий и адресат находятся в разных местах) и, тем более, в гипотаксисе и нарративе «дейктические полномочия» передаются другому лицу.

Во время работы над «Новым большим немецко-русским словарем» возникла проблема при попытке адекватно представить ряд дейктических слов. В немецком языке есть огромное количество слов с элемантами *her* (≈ ‘сюда’) и *hin* (≈ ‘туда’), выражающими идею направления движения. Слова *hin* и *her* употребляются также и самостоятельно в различных комбинациях с глаголами и адвербиальными конструкциями для выражения пространственных отношений между участниками ситуации.

Традиционно семантика этих дейктических элементов описывалась как ‘по направлению **к** говорящему’ для *her* и ‘по направлению **от** говорящего’ для *hin*. Например: *komm her!* ‘иди сюда!’; *Bier her!* ‘пиво давай <сюда>!’; с другой стороны, *geh hin!* ‘иди туда!’; *nach oben hin* ‘вверх (при том, что говорящий находится внизу)’.

Казалось бы, эти формулировки могут быть использованы в качестве комментария и в лексикографии – что реально и имеет место во всех известных нам толковых немецких словарях. Особенно необходим подобный комментарий в немецко-русском словаре. Дело в том, что в русском языке нет таких системно организованных (т.е. почти грамматических) средств для выражения этих смыслов. Так, глаголы *hereinführen* и *hineinführen* оба переводятся на русский язык с помощью глаголов *вводить* / *ввозить*, так что идея ‘вводить/ввозить сюда, т.е. по направлению к говорящему’ vs. ‘вводить/ввозить туда, т.е. по направлению от говорящего’ в русском языке часто вообще не выражается.

Однако более внимательный анализ показал, что комментарии к немецким *her* и *hin*, апеллирующие к традиционному правилу, искажают реальную картину. Так, человек, желающий войти в некоторое помещение (в стандартной ситуации) спросит, постучав, *Darf ich herein?* (т.е. *Можно мне войти?*, сокращенная форма от *Darf ich hereinkommen?* ‘Можно мне сюда войти?’), а не *Darf ich hinein?* (букв. ≈ ‘Можно мне туда войти?’), т.е. явно нарушит традиционное правило: предполагаемое перемещение в пространстве должно совершаться не по направлению **к** говорящему, а **от** его исходного местонахождения; а именно, по направлению к адресату.

Предпочтительность формулы *Darf ich herein?* подтверждается данными немецкоязычного Интернета. На *Darf ich herein?* нами обнаружено более 600 контекстов употребления, в то время как форма *Darf ich hinein?* представлена только 50 контекстами, большинство из которых не вполне стандартны или подразумевают принципиально иное пространственное расположение коммуникантов: вопрос *Darf ich hinein?* (сокращение от *Darf ich hineingehen?* ‘Можно мне туда войти?’) обращен не к адресату, находящемуся внутри помещения, в которое хочет войти говорящий, а к адресату, стоящему рядом с говорящим снаружи.

Можно думать, что мы имеем здесь дело с достаточно общим явлением, которое можно описать так. Вопрос *Darf ich hinein?* уместен в канонической (см. Lyons 1979: 579) коммуникативной ситуации, когда говорящий и адресат находятся в **одном и том же** месте. А вопрос *Darf ich herein?* задается в неканонической (хотя и речевой) ситуации – когда говорящий и адресат находятся в **разных** местах. Тем самым говорящий оказывается перед

021-Dobrovolskiy-Paducheva:_Layout 1 18.05.2008 14:44 Page 141

выбором – на кого ему ориентироваться, на себя или на адресата. Разные языки закрепили разные ориентации; немецкий выбрал ориентацию на адресата (так сказать, «политкорректную»)[1]. Так что преобладание вопроса с *herein* обусловлено естественным количественным преобладанием такого внешнего контекста для этого вопроса, когда говорящий и адресат находятся по разные стороны от двери, т.е. в разных пространствах.

Надо сказать, что русский говорящий был бы в затруднении, если бы захотел в этой ситуации употребить местоименное наречие: *Можно мне войти туда?* – это нонсенс, но и *Можно мне сюда войти?* не лучше. Полнейшее «ни туда, ни сюда».

Примеры, близкие к немецкому, можно найти в китайском языке: дейктические элементы с аналогичной функцией (*lái* ≈ ‘к Х-у, сюда’ и *qù* ≈ ‘от Х-а, туда’) ведут себя в соответствующих ситуациях так же, как в примере «*Darf ich herein?*», когда говорящий мысленно ставит себя в позицию адресата: *Kě yǐ jìn lái ma?* (букв. ≈ ‘Можно войти сюда?’). Так что можно предположить, что мы здесь имеем дело не с уникальной особенностью немецкого языка, а с более или менее универсальным явлением, связанным с общими принципами функционирования дейксиса. Видимо, в такой ситуации говорящий всегда может встать на позицию адресата (как бы предвосхитив его последующую реплику), и разница между языками состоит лишь в степени обязательности этого сдвига позиции. В самом деле, в случае грамматикализации дейктических элементов язык должен обеспечивать говорящему какой-то выход из положения.

Итак, субъектом дейксиса может быть не только говорящий.

Универсальность принципов перехода от одной ориентации дейктических элементов на другую подтверждается тем, что и в русском языке, хотя он и не располагает подобными регулярными лексическими показателями пространственного дейксиса, обнаруживается немало аналогичных стратегий употребления дейктических слов. Ср., например, употребление русских наречий *справа* и *слева* в контекстах типа

(1) Ты не видел мои очки? – Да вот же они у тебя, слева.

При этом *слева* с точки зрения говорящего может оказаться и ‘справа’. Всё дело в том, принял ли говорящий точку зрения своего адресата-визави или остался на своей.

Очевидно, в немецком то же самое: когда говорящий спрашивает *Darf ich herein?*, он встает на точку зрения адресата. Разница в том, что в русской ситуации *слева-справа* у говорящего есть выбор, а в случае немецкого *Darf ich herein?* выбор предопределен узусом. Такой же переход на точку зрения адресата в латинских так называемых «эпистолярных временах», суть которых, согласно Есперсен 1958: 343, состоит в том, что автор письма переносится в то время, когда это письмо будет читаться, и поэтому употребляет имперфект или перфект, где для нас единственно естественной является форма настоящего времени.

Отсюда можно сделать вывод, что более точным описанием дейктических слов, подобных немецким *her* и *hin*, должен быть комментарий типа «по направлению к говорящему – или к адресату, если говорящий принимает его точку зрения» vs. «по направлению от говорящего – или от адресата, если говорящий принимает его точку зрения».

Однако и включение адресата в субъекты дейксиса не покрывает всех релевантных случаев. Так, в нарративе субъектом дейксиса может оказаться не говорящий и не адресат, а более или менее любой персонаж, которого автор наделяет функцией субъекта восприятия: любое лицо может быть «заместителем» говорящего. Например, предложение *er trat herein* ‘он вошел’ предполагает наблюдателя, который воспринимает ситуацию не с точки зрения субъекта действия (*er*), а как бы находясь в помещении, в которое вошел субъект. Русским предложением *он вошел* может переводиться и *er trat herein*, и *er trat hinein*. В последнем случае наблюдатель находится снаружи.

В русском языке есть способ выразить положение наблюдателя – это порядок слов и коммуникативная перспектива предложения. Так, в (2б), практически однозначно, наблюдатель в кафе, а в (2а) – снаружи:

(2) а. Полицейский вошел в кафе;

б. В кафе вошел полицейский.

Итак, при объяснении значения дейктических элементов *her* и *hin* более адекватным представляется комментарий, в котором «адресат» будет заменен на «лицо, точку зрения которого принимает говорящий», т.е. должно быть что-то вроде «по направлению к говорящему или лицу, точку зрения которого принимает говорящий» vs. «по направлению от говорящего или лица, точку зрения которого принимает говорящий». Только в этом случае будут приняты во внимание также и контексты нарративного употребления данных дейктических слов. Это «лицо», являющееся нарративным заместителем говорящего, естественно назвать наблюдателем.

[1] Выбор между *hin-* и *her*-формами доставляет, видимо, определенные трудности и носителям немецкого языка, поэтому в разговорном языке почти всегда употребляются редуцированные формы типа *rein*, в которых снято интересующее нас противопоставление (ср., например, Krause 1998). Так, *rein* – это разговорный аналог как *herein*, так и *hinein*. Иными словами, придерживающийся разговорного стандарта человек, желающий войти в некоторое помещение, спросит *Darf ich rein?*, оставив тем самым открытым вопрос о выборе ориентации.

021-Dobrovolskiy-Paducheva:_Layout 1 18.05.2008 14:44 Page 142

Иными словами, в речевом режиме соперником говорящему может быть только адресат; а в нарративе это может быть и «третье лицо».

Замечание. Интересно в этой связи слово *сцена*, в контексте которого глаголы движения по-разному употребляются в разных языках. По-французски нельзя сказать *sortir sur scene*, что соответствовало бы пословно русскому *выйти на сцену*: нужно сказать *entrer sur scene*, букв. ‘войти на сцену’. Но здесь дело не в позиции наблюдателя (ведь и войти и выйти можно как **ко** мне, так и **от** меня), а в том, что в русской модели мира сцена мыслиться как **открытое** пространство, на которое выходят из какого-то более **замкнутого** (см. Апресян 1995: 490), а для французского свойства исходного пространства не играют роли. Иными словами, русский глагол сохраняет идею предварительного пребывания субъекта в замкнутом пространстве (возможно, темном или плохом), а французский выражает только идею ‘появиться’ (т.е. ‘начать находиться’). Прошлое местонахождение субъекта французский выбор глагола никак не отражает.

В ряде конструкций употребление *her* или *hin* лексикализовано. Т.е. в этих конструкциях у говорящего нет свободы в выборе эгоцентрика. Можно подумать, что подобная лексикализация объясняется исключительно капризами узуса. Однако на самом деле выбор *her* или *hin* в таких конструкциях семантически мотивирован. Дело в том, что в определенных ситуациях одна из возможных позиций наблюдателя оказывается более естественной, чем остальные. Например, если некто толкает перед собой тележку, то в этой ситуации можно сказать лишь *vor sich her schieben* (а не **vor sich hin schieben*). Вопрос, почему. В данной ситуации нет ни говорящего, ни адресата. Это нарратив. Понятно, что говорящего замещает некоторое другое лицо: повествователь или Erzählmedium, «через которого» описывается происходящее, т.е. глазами которого «пользуется» автор. Казалось бы, можно предположить, что в рассматриваемой ситуации в качестве Erzählmedium’а может выступать в том числе и субъект описываемого действия. В этом случае единственно уместным было бы выражение *vor sich hin schieben*; между тем, оно противоречит устоявшимся нормам. Видимо, в таких ситуациях более естественно встать на позицию стороннего наблюдателя, стоящего лицом к субъекту действия, т.е. наблюдателя, для которого перемещение в пространстве будет осуществляться в направлении ‘сюда’, а не ‘туда’.

Наши наблюдения за употреблением данных дейктических частиц в конструкциях с *vor sich* показали, что эти конструкции в высокой степени лексикализованы, причем эта фиксация зависит от естественности той или иной перспективы, т.е. позиции наблюдателя – что в свою очередь зависит от семантики соответствующего глагола. Так, ментальные глаголы, глаголы говорения и внутренних состояний, а также глаголы со значением физического действия, передающего некоторое внутреннее состояние, сочетаются с конструкцией *vor sich hin*. Примеры.

(3) а. Er grübelte *vor sich hin* und begann langsam zu verzweifeln.
Он ломал себе голову и постепенно начал приходить в отчаяние (букв. «размышлял/раздумывал про себя»).

б. Nun lief er weiter und murmelte immer *vor sich hin* <...>.
Он побежал дальше и все бормотал себе под нос <...> («про себя», «сам себе»).

в. Peter Smith, Journalist, fluchte leise *vor sich hin*.
Журналист Петер Смит тихо выругался про себя.

г. Die Schlafkrankheit hat ihren Namen dadurch erhalten, dass die Kranken *vor sich hin* dämmern.
Сонная болезнь, потому так и называется, что больные находятся как бы в полузабытьи (букв. ≈ «смеркаются про себя»).

д. Er nickte *vor sich hin* und sagte abschließend: «Das ist wichtig».
Он кивнул как бы сам себе и сказал подытоживая: «Это важно».

Как видно из примеров, конструкция *vor sich hin* приблизительно соответствует русским *про себя, сам себе, сам с собою*. Она фокусирует состояние субъекта и как бы задает перспективу «по направлению к субъекту», а не «от него». Соответственно, с точки зрения наблюдателя (а в контекстах 3 лица естественно предположить, что наблюдатель – это не сам субъект) перспектива переворачивается, т.е. с субъектом происходит нечто, направленное на него самого, а значит, «от наблюдателя». Видимо, по этой причине в составе конструкции закрепился дейктический элемент *hin* (‘от субъекта восприятия’), а не *her*. Это особенно хорошо видно на примере вербальных действий. В обычном случае вербальные действия направлены вовне. Именно так были бы прочитаны контексты (3б) и (3в), если бы глагол не сопровождался конструкцией *vor sich hin*. Функция этой конструкции – указать, что действие субъекта направлено не на адресата, а на самого себя. Ср. русские *выругался* и *выругался про себя*. Дейктический элемент *hin* осмыслен здесь лишь при допущении, что субъектом дейксиса является наблюдатель, а не лицо, выраженное подлежащим.

021-Dobrovolskiy-Paducheva:_Layout 1 18.05.2008 14:44 Page 143

Рассмотрим теперь функционирование конструкции *vor sich her*.

(4) а. Das Pferd wäre auch durchgegangen, wenn es nicht ein Bauer aufgehalten hätte, der <...> eine Kuh *vor sich her* trieb.

Лошадь тоже бы прошла, если бы ее не задержал крестянин <...>, гнавший перед собой корову.

б. Während Christine nun diesen Wagen *vor sich her* schiebt <...>.

Пока Кристина толкает перед собой машину <...>.

в. Blogs <...> tragen keine Ziele *vor sich her* <...>.

Блоги <...> не ориентированы на какие-либо цели <...> (букв. «не несут перед собой никаких целей», т.е. целей, направленных на что-л./кого-л.).

г. Die können 30 Jahre einfach so *vor sich her* unterrichten.

Они [речь идет о плохих учителях] могут 30 лет подряд просто так тихо себе преподавать.

Конструкция *vor sich her* отличается от *vor sich hin* тем, что описываемое действие (чаще всего физическое) направлено вовне. Соответственно, с точки зрения внешнего наблюдателя действие совершается в направлении «к нему», а не «от него», что и выражатся словом *her*.

Очевидно, что подобные наблюдения весьма значимы для лексикографии и контрастивной лингвистики (а возможно, и для общей теории дейксиса).

Еще один сложный момент в употреблении рассматриваемых дейктических слов – функционирование вопросительных местоименных наречий, в состав которых входят *her* или *hin*. Значение данных дейктических элементов (и соответственно их перевод на русский язык) меняется в зависимости от того, употребляется ли это наречие в прямом вопросе или в качестве относительного наречия в гипотаксисе. Это связано с тем, что в речевом акте вопроса субъектом дейксиса является говорящий, а в гипотаксисе – субъект матричного предложения. Ср., например:

(5) а. *Woherunter* sind sie gefahren?

Это по какой такой дороге они поехали вниз? [скорее всего, в контексте переспроса с эмфатическим ударением – когда важнее место, чем направление; в нефорсированных вопросительных контекстах форма *woherunter* и ей подобные воспринимаются как явно устаревшие; сейчас в этом случае скажут *Wo sind sie heruntergefahren?* или даже скорее *Wo sind sie runtergefahren?*]

б. Alle schauten entsetzt auf den Felsen, *woherunter* der Bergsteiger gestürzt war.

Все в ужасе посмотрели на скалу, откуда <с которой> сорвался альпинист.

Выбор наречия *woherunter* (в отличие от *wohinunter*) мотивирован в случае (5а) пространственным положением говорящего. Понятно, что в момент речи говорящий находится внизу (если бы он стоял на горе, он спросил бы *wohinunter...?*). А в случае (5б) выбор наречия объясняется местонахождением людей, смотрящих на скалу (*alle*). Контекст ясно показывает, что они стоят внизу и смотрят на скалу снизу вверх.

Наши *hin-* и *her*-слова показывают (как им и положено), где наблюдатель. *Woherunter* – это как бы «где-сюда-низ», а *wohinunter* – «где-туда-низ». Соответственно, если наблюдатель наверху, он скажет «где-туда-низ», т.е. *wohinunter*, по-русски что-то вроде *куда (вниз*), а если он внизу – *woherunter* – «где-сюда-низ», т.е. что-то вроде *откуда (вниз)*.

Что стоит за всеми этими явлениями по сути? Очевидно, речь идет об общих проблемах различия в употреблении и интерпретации дейксиса в зависимости от того или иного режима функционирования языка. Иными словами, для понимания функционирования подобных дейктических элементов необходимо развести три различных режима их употребления: (а) диалоговый (речевой) режим, (б) гипотаксис, (в) нарратив.

Известно, что повествовательный текст (нарратив) функционирует в обедненном контексте. Восприятие и интерпретация нарратива происходит в условиях особой – редуцированной, ущербной коммуникативной ситуации: в контексте, где нет полноценного говорящего: автор и читатель не связаны единством места и времени, не имеют общего поля зрения и не могут видеть жестов друг друга. А если речь идет о fiction, то они вообще не принадлежат МИРУ текста.

Двум разным контекстам употребления языка соответствуют разные РЕЖИМЫ ИНТЕРПРЕТАЦИИ текста: различаются РЕЧЕВОЙ режим и НАРРАТИВНЫЙ.

Говорящий может присутствовать в семантике языковых единиц в разных ипостасях. Прежде всего – как субъект РЕЧИ. А кроме того, говорящий может быть представлен в тексте как субъект ДЕЙКСИСА, субъект ВОСПРИЯТИЯ и субъект СОЗНАНИЯ (т.е. модальности, мнения, оценки, эмоции, и тому подобное).

Эгоцентриками являются грамматические категории (такие, как время и вид); слова типа *я, ты, здесь, сейчас, тут, там, этот, тот*, *вон*, *вот*; типа *к счастью*, *все-таки*, *по правде говоря*; *да, нет, в самом деле* и т.д. Если во фразе есть эгоцентрик, это значит, что в обозначаемую ситуацию включен СУБЪЕКТ – который в поверхностной структуре может быть и не выражен.

021-Dobrovolskiy-Paducheva:_Layout 1 18.05.2008 14:44 Page 144

В первом приближении этим субъектом является говорящий (например, *увы* и *к счастью* выражают сожаление и радость говорящего). Однако если приглядеться, то оказывается, что одни эгоцентрики действительно предполагают **именно** говорящего, а для других это может быть как говорящий, так и другое лицо. Так что эгоцентрики бывают ПЕРВИЧНЫЕ и ВТОРИЧНЫЕ. Первичные эгоцентрики реализуют свой смысл только в условиях канонической речевой ситуации и ориентируются всегда **только** на **полноценного** говорящего; а вторичные могут быть ориентированы не только на говорящего, но также и на другое лицо. Так, в примере (6) *какую-то* – вторичный эгоцентрик; предполагает субъекта неопределенности-незнания, и в (6а) этим субъектом является говорящий, а в (6б) – подлежащее 3 лица, Маша:

(6) а. Она хочет спеть *какую-то* песню = ‘я не знаю, какую’ [*какую-то* ориентировано на говорящего, который является субъектом незнания];

б. Маша сказала, что на столе *какая-то* записка [субъект незнания – Маша].

Можно сказать, что у слова *какой-то* есть семантическая валентность на субъект незнания, и, поскольку *какой-то* – вторичный эгоцентрик, субъект незнания допускает интерпретацию в разных режимах.

Вторичными эгоцентриками являются также и рассматриваемые нами немецкие *her* и *hin*. Мы покажем, что *her* и *hin*, будучи вторичными эгоцентриками, допускают три интерпретации:

(i) интерпретация в речевом режиме, субъект дейксиса – говорящий;

(ii) интерпретация в гипотаксическом режиме, субъект дейксиса – подлежащее матричного предложения;

(iii) интерпретация в нарративном режиме, субъект дейксиса – какое-то лицо в контексте.

Ниже приводятся примеры из параллельного корпуса НКРЯ. В качестве источника мы взяли сказку Э.Т.А. Гофмана «Крошка Цахес»: в оригинате (E.T.A. Hoffmann «Klein Zaches genannt Zinnober») и в русском переводе.

(i) интерпретация в речевом режиме, субъект дейксиса – говорящий:

(7) «Gott im Himmel», rief der Kammerdiener entsetzt, «aus dem Fenster der gnädigen Exzellenz kuckte ja das kleine abscheuliche Ungetüm *heraus*. – Was ist das?»
– Боже праведный, – вскричал камердинер в ужасе. – Да ведь это мерзкое чудище выглянуло из окна их превосходительства. – Что б это значило?

Здесь говорящий явно находится снаружи, поэтому с его точки зрения возможно только *heraus*. *Hinaus* было бы неверно. Работает то самое стандартное правило, которое записано в словарях и учебниках. Наиболее точно его можно записать следующим образом: *heraus* означает ‘наружу (при обозначении перемещения из более замкнутого в более открытое пространство по направлению **к** говорящему)’. *Hinaus* означает ‘наружу (при обозначении перемещения из более замкнутого в более открытое пространство по направлению **от** говорящего)’. Здесь, в отличие от русского языка, противопоставляется не ВНУТРЬ и НАРУЖУ, а ОТ и К: *hinaus / heraus* – оба НАРУЖУ (ВНУТРЬ было бы *hinein / herein*.) Разница в том, что одно НАРУЖУ ОТ, а другое – НАРУЖУ К.

(ii) интерпретация в гипотаксическом контексте, субъект дейксиса – подлежащее матричного предложения:

(8) Er hatte sich geirrt, denn aus dem Gebüsch heraustretend, gewahrte er ganz in der Ferne, wie noch ein anderer stattlicher Reiter sich zu dem Kleinen gesellte und wie nun beide in das Tor von Kerepes *hinein*ritten.
Но он ошибся. Выйдя на опушку, он увидел, как вдалеке к малышу присоединился другой всадник, статный с виду, и оба уже въезжали в ворота Керепеса.

Здесь субъектом дейксиса является субъект главного предложения (т.е. *er*), который, стоя в лесу, наблюдает, как всадники въезжают в ворота. С точки зрения наблюдателя это перемещение из более открытого пространство в более замкнутое по направлению от него, наблюдателя. Дейктический компонент *hinein* в составе глагольной формы *hineinritten* показывает (в силу общих законов интерпретации вторичных дейктических элементов в гипотаксическом контексте), что в качестве наблюдателя выбран именно данный персонаж (а не, например, кто-либо, находящиеся внутри Керепеса). К тому же этот персонаж является просто субъектом наблюдения (*он увидел, как*).

(iii) интерпретация в нарративном режиме, субъект дейксиса – какое-то лицо в контексте:

(9) Prosper Alpanus senkte sich *herab* zu dem Jüngling, an dessen Seite er Platz nahm, während die Libelle aufflog in ein Gebüsch und in den Gesang einstimmte, der durch den ganzen Wald tönte.
Проспер Альпанус спустился к юноше и сел подле него, стрекоза упорхнула в кусты, вторя пению, наполнявшему весь лес.

021-Dobrovolskiy-Paducheva:_Layout 1 18.05.2008 14:44 Page 145

Здесь в принципе возможно и *hinab*. Но в этом случае субъектом дейксиса был бы Проспер Альпанус. А в перспективе повествования, выбранной автором, субъект дейксиса – юноша. Как видно из перевода, русский язык не располагает стандартными средствами, позволяющими выразить эти различия: ‘спустился сюда’ vs. ‘спустился туда’. Более точный перевод был бы «Проспер Альпанус спустился сюда к юноше».

Наоборот, *hinauf* в предложении (10) – которое в тексте предшествует предложению (9) – развертывает событие в противоположной перспективе. Субъект восприятия здесь тот же юноша. Но в одном случае движение направлено к нему, потому что Проспер Альпанус спустился к нему вниз, а в другом – от него (ср. движение глаз юноши, когда он посмотрел наверх).

(10) Er schaute *hinauf* und erblickte staunend Prosper Alpanus, der auf einem wunderbaren Insekt, das einer in den herrlichsten Farben prunkenden Libelle nicht unähnlich, daherschwebte.
Он поднял глаза и с изумлением увидел Проспера Альпануса, летевшего к нему на каком-то диковинном насекомом, не лишенном сходства с великолепной, сверкающей всеми красками стрекозой.

Он (юноша) – субъект восприятия и субъект дейксиса. Он находится внизу, а Проспер Альпанус на стрекозе – наверху. Следовательно, адвербиал *herauf* (≈ ‘сюда наверх’) здесь был бы неуместен. С точки зрения юноши, он смотрит ‘туда наверх’, т.е. *hinauf*.

Итак, мы видим, что *her* и *hin* могут, как и другие вторичные эгоцентрики, быть ориентированы не на говорящего.

Интересно рассмотреть еще одно явление, которое можно трактовать как отказ говорящего от своих дейктических прав и переход на точку зрения адресата. В стихотворении Бродского (пример из Падучева 1996: 264) местоимение 2 лица *ты*, фактически, обозначает говорящего:

(11) Пространство в тысячу ли…
Тысяча означает, что ты сейчас вдали
От родимого крова…

И.И.Ковтунова предлагает рассматривать этот сдвиг от 1-го лица ко 2-му как происходящий в два приема. На первом этапе говорящий отказывается от своих прав на называние себя в 1-м лице и переходит на положение 3-го лица, как в нарративе. «Речь о себе в третьем лице дает <…> возможность сделать себя объектом восприятия со стороны адресата речи – друга, возлюбленной, вечного лирического *ты*, вызывает к жизни и вводит ее <может быть, ее /его?> точку зрения. Речь от первого лица этой возможности не дает, поскольку автор и объект речи слиты в одном лице. Третье лицо их разъединяет, устанавливает между ними дистанцию.» (Ковтунова 1986: 94).

В качестве примера перехода говорящего на положение 3-го лица И.И.Ковтунова приводит 3 лицо лирического героя в стихах Блока из цикла «Кармен». Для нарратива на переходе от 1 лица к 3-му все заканчивается. Между тем в лирическом стихотворении это 3 лицо говорящего может естественно стать вторым для изначально заложенного в структуру лирического жанра внешнего адресата. На этом пути мы получаем объяснение для *ты* говорящего в строчках из Бродского. Оказывается, что это *ты* (от которого два шага до обобщенно-личного *ты*, как оно описано в Булыгина 1990) достигается тем же переходом говорящего на позицию адресата, который дает *herein* в нашем исходном примере *Darf ich herein?*

И последний пример. Понятие режима интерпретации позволяет дать теоретическое обоснование истокам неоднозначности, связанной с противопоставлением внешнего и внутреннего адресата в лирическом стихотворении, см. Виноградов об Ахматовой (Виноградов 1935/1976) и, позднее, Ковтунова 1986. Мы продемонстрируем эту неоднозначность на примере стихотворения Пушкина «Город пышный, город бедный… 1828» (на это стихотворение, точнее – на возможность неоднозначного его понимания – обратила наше внимание Е.А.Гришина).

(12) Город пышный, город бедный,
Дух неволи, стройный вид,
Свод небес зелено-бледный,
Скука, холод и гранит –
Все же мне вас жаль немножко,
Потому что здесь порой
Ходит маленькая ножка,
Вьется локон золотой.

Вне контекста стихотворение остается не вполне понятным; неясен смысл по крайней мере трех слов (два из них – дейктические): *вас, жаль* и *здесь*.

Одно понимание – в монологическом режиме, при котором *вас* обращено к атрибутам Петербурга[2], т.е.

[2] Обращение к неодушевленным объектам в поэзии вполне обычно, ср. хотя бы: Что будет со мною, старинные плиты? (Пастернак. «Марбург») Большой подбор примеров – в Ковтунова 1986.

021-Dobrovolskiy-Paducheva:_Layout 1 18.05.2008 14:44 Page 146

является внутренним адресатом. В этом случае *здесь* должно интерпретироваться анафорически – как относящееся к ранее косвенно упомянутому Петербургу и всему неодушевленному, что с ним метонимически связано. Непонятно, правда, каким образом эти неодушевленности могут вызывать сожаление.

Другое понимание возникает при погружении высказывания в диалогический контекст: оно может быть обращено к адресату, находящемуся в другом пространстве. В этом случае *здесь* приобретает чисто дейктическую интерпретацию и противопоставлено тому *там*, где находится адресат (единоличное право говорящего на *здесь* – в отличие от права на *там*, которое, как мы видели, говорящий иногда должен делить с адресатом, – в русском языке неотъемлемо). Естественную интерпретацию получает *жаль*: говорящий жалеет адресата, который находится в другом пространстве и не может получать наслаждения от ножки и локона.

Короткий комментарий Б.В. Томашеского («Написано в связи с готовящимся отъездом поэта из Петербурга в Михайловское. Последние две строки относятся к А.А. Олениной (1808-1888)») ставит всё на свои места. Никакого внешнего адресата нет. Адресат внутренний; *здесь* анафорическое; всё дело в особом значении слова *жаль*, которое должно пониматься как ‘жаль расставаться’, см. о семантике сожаления Зализняк 1988. Так что неоднозначности того типа, за которым охотился В.В. Виноградов (1976), здесь, пожалуй, нет (т.е. Пушкин не имел ничего такого в виду).

И.И. Ковтунова (с. 118) отмечает в этом стихотворении неоднозначность, которую нельзя не признать: первая строфа может пониматься как номинативное предложение, тогда как *вас* в значении внутреннего адресата требует, чтобы она была обращением.

В заключение остается повторить, что обращение к коммуникативной ситуации позволяет более точно описать значение и особенности употребления ряда конкретных дейктических элементов и, тем самым, глубже проникнуть в суть семантики дейксиса.

Список литературы

1. Апресян Ю.Д. Избранные труды. Т. 2. М.: Языки рус. культуры, 1995.

2. Булыгина Т.В. Я, ты и другие в русской грамматике // Res philologica. Филологические исследования. Памяти акад. Г.В. Степанова. М.-Л.: Наука, 1990, с. 111-126.

3. Виноградов В.В. Избранные труды. Поэтика русской литературы. М.: Наука, 1976.

4. Есперсен О. Философия грамматики. М.: Изд-во иностр. лит., 1958. (Англ. ориг.: Jespersen O. The Philosophy of Grammar. London, 1924).

5. Зализняк Анна А. О семантике сожаления // Логический анализ языка. Прагматика и проблемы интенсиональности. М.: Наука, 1988. С.189-213.

6. Ковтунова И.И. Поэтический синтаксис. М.: Наука, 1986.

7. Падучева Е.В. Семантические исследования: Семантика времени и вида в русском языке. Семантика нарратива. М.: Языки рус. культуры, 1996.

8. Krause M. Überlegungen zu *hin-/her-* + Präposition // Particulae particularum. Tübingen: Stauffenburg, 1998. S.195-217.

9. Lyons J. Semantics. Vol. 1–2. London etc.: Cambridge Univ. Press, 1977.

022-Drujkin_Cinman:_Layout 1 18.05.2008 14:50 Page 147

СИНТАКСИЧЕСКИЙ АНАЛИЗАТОР ЛИНГВИСТИЧЕСКОГО ПРОЦЕССОРА ЭТАП-3: ЭКСПЕРИМЕНТЫ ПО РАНЖИРОВАНИЮ СИНТАКСИЧЕСКИХ ГИПОТЕЗ[1]

THE PARSER OF ETAP-3 LINGUISTIC PROCESSOR: EXPERIMENTS ON RANKING SYNTACTIC HYPOTHESES

Дружкин К.Ю*. (druzhkin@iitp.ru),* ***Цинман Л.Л.****(cinman@iitp.ru)*
ИППИ РАН

Предпринимается попытка оптимизировать работу синтаксического анализатора в лингвистическом процессоре ЭТАП-3. Предлагается изменить описание синтаксических правил, чтобы возникающие при анализе гипотезы ранжировались по вероятности вхождения в результирующую древесную синтаксическую структуру фразы. Приводятся результаты эксперимента.

0. Вводные замечания

В предлагаемой статье речь пойдет о русском синтаксическом анализаторе многоцелевого лингвистического процессора ЭТАП-3, разработанном в лаборатории компьютерной лингвистики Института проблем передачи информации РАН (см. об этом процессоре и его приложениях, в частности, [1] – [4]).

В первом разделе кратко описывается устройство синтаксического анализатора в системе ЭТАП-3 и объясняется, почему полнота описания приводит к порождению большого числа неверных гипотез. Во втором разделе обсуждается один возможный способ решения этой проблемы. Третий раздел содержит описание проведённого эксперимента и его результатов.

1.Проблема: перепорождение гипотез
1.1. Синтаксический анализ в системе ЭТАП

Синтаксическая структура фразы в виде дерева зависимостей строится в лингвистическом процессоре ЭТАП-3 с помощью специальных правил (синтагм). Этих правил для каждого из рабочих языков системы (русского и английского) насчитывается несколько сотен. Все они бинарны: это означает, что любая синтагма позволяет связать некоторым синтаксическим отношением два слова фразы, если все условия этой синтагмы, описывающие контекст данной пары слов во фразе, выполнены. Говоря чуть более строго, синтагма связывает синтаксическим отношением не слова фразы, а некоторую пару омонимов[2] этих слов, если они представлены в начале синтаксического анализа несколькими (морфологическими и/или лексическими) омонимами. Таким образом, омонимы слов фразы могут связываться синтаксическими отношениями независимо друг от друга.

В результате работы синтагм на первом этапе синтаксического анализа возникает граф гипотетических синтаксических связей (синтаксических гипотез). На дальнейших этапах синтаксического анализатора посторонние связи различными средствами отфильтровываются, и из графа синтаксических гипотез выделяется дерево синтаксической структуры фразы. Иными словами, в основе алгоритма синтаксического анализа системы ЭТАП-3 лежит так называемый “фильтровый метод”.

Успех работы синтаксического анализатора такого типа зависит от решения двух задач.

[1] Настоящая работа частично поддержана Российским фондом фундаментальных исследований (грант № 07-06-00339), которому авторы выражают признательность.

[2] Термин «омоним» применяется здесь не совсем так, как в традиционной лексикографии: под омонимом некоторого слова (точнее, словоформы) понимается его конкретная морфологическая и/или лексическая интерпретация. Например, словоформа течь может интерпретироваться 1) как глагол в инфинитиве (ср. вода перестала течь), 2) как существительное в именительном падеже единственного числа (ср. течь была устранена) и 3) как существительное в винительном падеже единственного числа (ср. надо устранить течь). Соответственно, словоформа течь имеет три омонима.

022-Drujkin_Cinman:_Layout 1 18.05.2008 14:50 Page 148

Первая задача: создание корпуса синтагм, максимально полно описывающих синтаксические явления в естественном языке.

Вторая задача: создание развитой совокупности фильтровых средств, позволяющих выделить из графа гипотетических связей искомую синтаксическую структуру.

1.2. Уроки массовой разметки текстов

Система ЭТАП-3 находится в экспериментальной эксплуатации уже довольно давно. В частности, в последние годы с помощью этой системы в рамках программы "Создание глубоко аннотированного корпуса русского языка" (подробнее о ней см. [3]) были синтаксически размечены десятки тысяч фраз из разного рода текстов (сейчас в корпусе около 37 000 фраз).

Все синтаксические структуры этих фраз сначала «начерно» строились системой ЭТАП-3, а затем вручную редактировались специалистами-лингвистами. Массовая синтаксическая разметка текстов стала важнейшей проверкой возможностей нашего синтаксического анализатора. Анализ этой работы выявил два важных обстоятельства.

Первое: корпус русских синтагм обладает достаточной полнотой.

Действительно, для абсолютного большинства фраз, слова которых представлены в словаре ЭТАПа-3, среди гипотетических связей графа, построенного на основании синтагм, присутствуют все связи, которые должны составить правильную синтаксическую структуру этой фразы.

Второе: корпус русских синтагм в определенной степени избыточен.

Это означает, в частности, что для ряда фраз, содержательно не являющихся омонимичными, из графа гипотетических связей можно выделить несколько различных деревьев, все синтаксические связи в которых удовлетворяют нашему описанию синтаксиса.

1.3. Пример посторонних интерпретаций

Приведем простой пример: пусть на вход синтаксического анализатора поступает вопросительное предложение *Что делает правительство?* С точки зрения любого человека – носителя русского языка, это предложение совершенно однозначно:. слово *правительство* здесь является субъектом, подлежащим, а слово *что* - прямым дополнением глагола *делает*. С точки же зрения нашего парсера это предложение допускает и другие интерпретации: в частности, 1) слово *что* может интерпретироваться как подлежащее, а *правительство* – как дополнение при глаголе *делает*; 2) слово *что* может интерпретироваться как союз, вводящий неполное предложение (такая интерпретация осмыслена, например, в контексте типа *Чем ты недоволен? Что ничего не делает президент? Что делает правительство?*). Очевидно, что вероятность того, что какая либо из этих двух последних интерпретаций адекватно отражает структуру нашего предложения в каком-либо тексте, исчезающее мала.

1.4. Полнота описания приводит к избыточности

Эти особенности лингвистического процессора системы ЭТАП (достаточная полнота и определенная избыточность), на наш взгляд, взаимосвязаны. Так, если лингвист, обслуживающий систему, встречает в тексте синтаксическую конструкцию, не учтенную в синтагмах, то ему достаточно подправить одну из соответствующих синтагм или создать новую, чтобы возникло недостающее синтаксическое отношение. Однако часто бывает так, что некоторая языковая конфигурация (скажем, последовательность словоформ, принадлежащих определенным лексико-грамматическим классам), будучи погружена в другие контексты, образует другую синтаксическую конструкцию и должна анализироваться уже иначе. Предусмотреть все эти контексты при написании синтагм, по-видимому, невозможно в принципе. Из этого следует, что синтагмы неизбежно будут порождать в ряде случаев лишние, неверные синтаксические гипотезы. Получается некоторый парадокс: чем полнее описан синтаксис языка, тем больше посторонних синтаксических гипотез будет возникать на первом этапе синтаксического анализа.

Как показывает опыт эксплуатации парсера ЭТАПа-3, для больших фраз количество гипотез может достигать величины 20-30 n, где n – число слов фразы. Естественно, что наряду с правильной синтаксической структурой из графа гипотетических связей могут быть выделены другие деревья зависимостей, которые, хотя и удовлетворяют всем условиям нашего синтаксического описания, но фактически представляют неправильные интерпретации фразы. Таким образом, стремление к полноте описания синтаксиса увеличивает избыточность (в указанном выше смысле) этого описания.

022-Drujkin_Cinman:_Layout 1 18.05.2008 14:50 Page 149

2. Возможное решение: ранжирование гипотез
2.1. Интерактивный выбор гипотез: достоинства и недостатки

У пользователя, работающего с системой ЭТАП-3, есть возможность затребовать альтернативный синтаксический разбор. В силу упомянутой выше полноты корпуса синтагм, при достаточно большом числе итераций для большинства фраз рано или поздно мы получим правильную структуру. Естественно, однако, что такая форма работы с системой неудобна и неприменима при работе с массовым материалом.

2.2. Ранжирование синтаксических гипотез

Важной проблемой для нас является оптимизация процесса выделения правильной синтаксической структуры из графа гипотетических связей. Необходимо стремиться к тому, чтобы правильная структура выделялась первой или одной из первых.

Определенные надежды в связи с этим мы возлагаем на ранжирование синтаксических гипотез, порождаемых синтагмами.

На первый взгляд может показаться, что предлагаемые ниже средства выполняют ту же роль, что внедрение в правиловую систему обучающего статистического компонента. На самом деле назначение этих средств принципиально иное: мы привносим в синтаксический анализатор ЭТАПа своего рода **эмпирическую статистику**, порожденную лингвистом-экспертом, который извлекает уроки из работы пусть несовершенной, но живой синтаксической системы и производит все более тонкую настройку этой системы. Этим достигаются две научные цели: 1) расширяются рамки возможностей построенной лингвистом действующей модели языка, и 2) точнее определяются границы этих возможностей. Кроме того, как мы надеемся показать ниже, с помощью подобных механизмов можно получить результаты, по мнению авторов, недостижимые с помощью методов чистой статистики, пусть даже с элементами машинного обучения.

Проиллюстрируем идею ранжирования синтаксических гипотез на примере всего одной, но весьма частотной русской синтагмы. Эта синтагма (в ЭТАПе она имеет имя *1-компл.11*) связывает 1-м комплетивным синтаксическим отношением переходный глагол (X) с прямым дополнением (Y). Она обслуживает, например, такие конструкции: (1) *купил* (X) *яблоко* (Y) – неосложненное прямое дополнение; (2) *купил* (X) *три яблока* (Y) – дополнение – количественная группа, «в целом» стоящая в винительном падеже (в которой существительное имеет родительный падеж); (3) *привезли* (X) *книг* (Y) *десять* – дополнение – аппроксимативно-количественная группа в винительном падеже (в которой также существительное стоит в родительном падеже); (4) *не купил* (X) *яблок* (Y) – дополнение в родительном падеже при глаголе с отрицанием; (5) *выпил* (X) *кваску* (Y), *выпил* (X) *пива* (Y) – дополнение в партитивном или родительном падеже с количественным значением.

Хотя в приведённых примерах дополнение выражено существительным, оно иногда может выражаться прилагательным, причастием или числительным. Далее, здесь дополнение располагается справа от глагола, но в ряде случаев может стоять и слева от него. Наконец, во всех рассмотренных фразах дополнение стоит рядом с глаголом-хозяином, но встречаются фразы, где оно расположено достаточно далеко от своего хозяина. Естественно, все эти возможности должны учитываться в синтагме.

Как следствие, если в большой фразе встретится переходный глагол, эта синтагма может породить большое число синтаксических гипотез, хотя известно, что глагол в принципе не может иметь более одного прямого дополнения.

Конечно, автор синтагмы, пытаясь сократить порождение посторонних гипотез, старается записать в условиях синтагмы всякого рода дополнительные сведения:

- условия согласования между X и Y (например, соблюдение семантических ограничений на заполнение соответствующей глагольной валентности);
- наличие или отсутствие в их ближайшем контексте некоторых классов слов;
- наличие или отсутствие в их ближайшем контексте тех или иных знаков препинания;
- необходимость или невозможность участия X и Y в некоторых других синтаксических связях, и т.п.

Но, поскольку синтагма должна быть рассчитана даже на крайне редко встречающиеся синтаксические конструкции (требование полноты!), то такого рода дополнительных ограничений можно сформулировать не так много.

Вернемся снова к синтагме *1-компл.11*, эскиз которой приведен выше. Рассмотрим гипотетическую фразу, в которой имеется переходный глагол X, справа рядом с ним существительное Y1 в винительном падеже, а слева от X, скажем, на расстоянии в 10 слов стоит числительное Y2, также в винительном падеже. Пусть оба претендента на роль прямого дополнения X, т.е. Y1 и Y2, прошли все необходимые проверки, требуемые синтагмой. Тогда наша синтагма породит две гипотетические связи с именем *1-компл*: X → Y1 и X → Y2. В то

022-Drujkin_Cinman:_Layout 1 18.05.2008 14:50 Page 150

же время очевидно, что вероятность вхождения в правильную синтаксическую структуру первой гипотезы неизмеримо выше, чем второй. Естественно, хотелось бы не потерять это знание и использовать его в предстоящем процессе фильтрации.

2.3. Ранжирование с помощью подправил

На подобных соображениях основано наше предложение ранжировать гипотезы, порождаемые синтагмами. На наш взгляд, было бы правильным по крайней мере основные синтагмы переписать в виде нескольких подправил по такой схеме: в первом подправиле записываются условия, при выполнении которых формируется гипотеза с пометой “сильная гипотеза”. анализатор переходит ко второму подправилу, выполнение условий в котором создает гипотезу с пометой “обычная гипотеза”. Если же условия и второго подправила не выполнены, то происходит переход к третьему подправилу, выполнение условий в котором создает гипотезу с пометой “слабая гипотеза”.

При таком описании синтагм синтаксические гипотезы в графе оказываются в значительной мере ранжированными. Мы рассчитываем на то, что если написать правила, которые в процессе фильтрации графа будут отдавать предпочтение “сильным гипотезам” перед “обычными гипотезами”, а “обычным гипотезам” – перед “слабыми гипотезами”, то мы достигнем правильной синтаксической структуры быстрее.

Отметим, что средства для выделения “слабых” и ”сильных” гипотез имеются и в действующем варианте синтаксического анализатора. “Сильные” гипотезы возникают, как правило, при описании устойчивых словосочетаний, а “слабыми” помечаются редко встречающиеся синтаксические конструкции. Однако эти средства используются лишь эпизодически и не связаны со специальной настройкой описания условий в синтагмах.

3. Эксперимент: создание «сильных синтагм»

Поскольку процедура переписывания синтагм по приведенной выше форме весьма трудоемка, мы посчитали разумным предварить такую деятельность некоторыми экспериментами, которые подтвердили бы ее целесообразность .

3.1. Содержание эксперимента

При проведении этих экспериментов мы поступили следующим образом. Чтобы не создавать помех работе синтаксического анализатора системы ЭТАП в основном режиме, для некоторых важных с нашей точки зрения русских синтагм были написаны синтагмы-“дублеры”. Условия в этих “дублерах” (сильных синтагмах) формулировались так, чтобы при их выполнении синтаксические связи, проводимые этими синтагмами, с высокой вероятностью вошли в синтаксическую структуру фразы (они помечались как “сильные гипотезы”). Был создан специальный режим работы синтаксического анализатора, при котором наряду с обычными синтагмами происходит обращение к синтагмам–дублерам. Так граф гипотетических связей дополнительно пополняется “сильными гипотезами” (в этом эксперименте созданием “слабых гипотез” мы не занимались). Было написано несколько правил, которые на разных этапах процесса фильтрации графа отдают предпочтение “сильным гипотезам”. В частности, на некотором этапе фильтрации “сильные гипотезы” стирают гипотезы, которые им противоречат.

Поскольку участие “сильных гипотез” в искомой синтаксической структуре фразы весьма вероятно, но все же не гарантировано, в экспериментальном режиме в случае аварийного завершения работы синтаксического анализатора (не построено никакой структуры) предусмотрен возврат к исходному графу гипотез и второй проход через процедуры фильтрации, но уже без учета “сильных гипотез”.

3.2. Пример сильной синтагмы

Вернёмся к синтагме, описывающей связь между переходным глаголом и его прямым дополнением. Сильный вариант этой синтагмы состоит из нескольких подправил (как и основная синтагма). Например, в первом подправиле, описывающем неосложненное прямое дополнение, рассматриваются только такие случаи, где дополнение (a) является существительным, (b) стоит в винительном падеже, (c) расположено справа от глагола, (d) не дальше, чем за два слова, (e) между ними нет существительных в винительном падеже, (f) между ними нет знаков препинания.

022-Drujkin_Cinman:_Layout 1 18.05.2008 14:50 Page 151

3.3. Результаты эксперимента

Результаты эксперимента оценивались следующим образом. Выше говорилось, что в нашем распоряжении имеется представительный корпус текстов, синтаксически размеченных ЭТАПом и отредактированных экспертами-лингвистами. Этот корпус естественно рассматривать как **эталон**, с которым следует сравнивать работу синтаксического анализатора ЭТАПа в его текущем состоянии. Была написана программа, которая автоматически сравнивает синтаксические структуры фраз эталона с синтаксическими структурами, построенными ЭТАПом, и на основе развитой системы штрафов оценивает работу ЭТАПа. С помощью этой программы мы смогли сравнить работу синтаксического анализатора ЭТАПа в двух режимах: штатном (использовавшемся в лаборатории на тот момент по умолчанию) и экспериментальном (с учетом “сильных гипотез”). Сравнение проводилось на эталонных текстах объемом около 1000 фраз.

Результат сравнения работы двух режимов синтаксического анализатора таков: у 90 фраз в экспериментальном режиме синтаксическая структура улучшилась (стала правильной или “ближе” к правильной), но для 19 фраз она стала хуже.

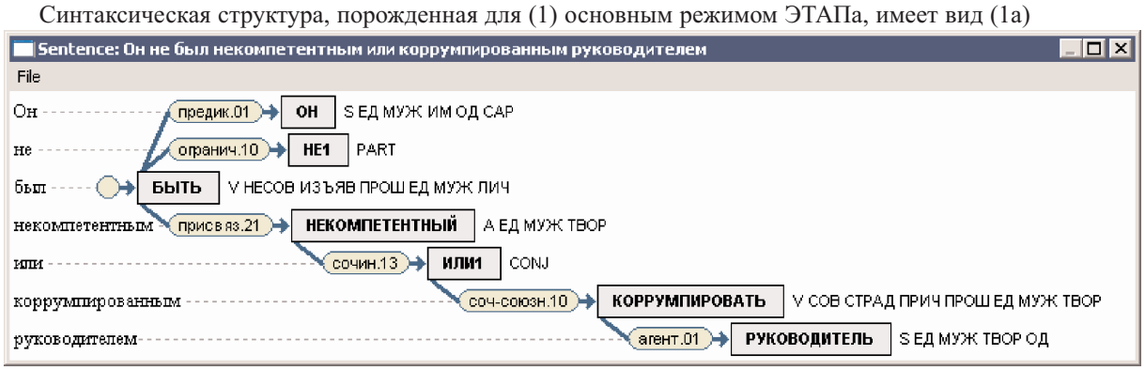
Эти цифры требуют комментариев. Напомним, что в эксперименте из общего числа 500 русских синтагм участвовало менее 20 (хотя и весьма важных) синтагм-дублеров “сильных синтагм”. Однако те “сильные синтагмы”, которые включены в эксперимент, следовало бы скорее назвать эскизами синтагм. Дело в том, что написание “сильной синтагмы” – весьма непростой процесс. Если сформулировать условия этой синтагмы достаточно жестко, то вероятность вхождения гипотезы, построенной этой синтагмой, в правильную синтаксическую структуру станет очень вероятной, но, с другой стороны, в реальных текстах условия такой синтагмы будут выполняться редко, а, значит, она редко будет порождать “сильные гипотезы”, и ранжирование в таком случае будет малопродуктивным. Если же условия необходимого контекста, описанного в синтагме, несколько ослабить, то синтагма чаще будет порождать “сильные гипотезы”, но вероятность их вхождения в искомую синтаксическую структуру уменьшится, что может привести к построению неправильной структуры. Разумный компромисс может быть установлен только в процессе масштабных экспериментов на больших текстах. Мы тренировали наши ”сильные синтагмы” на текстах с меньшим объемом (около 600 фраз). Для этих текстов результат сравнения двух режимов иной: для 51 фразы эксперимент улучшил структуру, для 3 фраз ухудшил.

3.4. Примеры работы синтаксического анализатора в основном и экспериментальном режимах

Мы проиллюстрируем различия в работе этих режимов на двух фразах.

(1) *Он не был некомпетентным или коррумпированным руководителем.*

Синтаксическая структура, порожденная для (1) основным режимом ЭТАПа, имеет вид (1а)

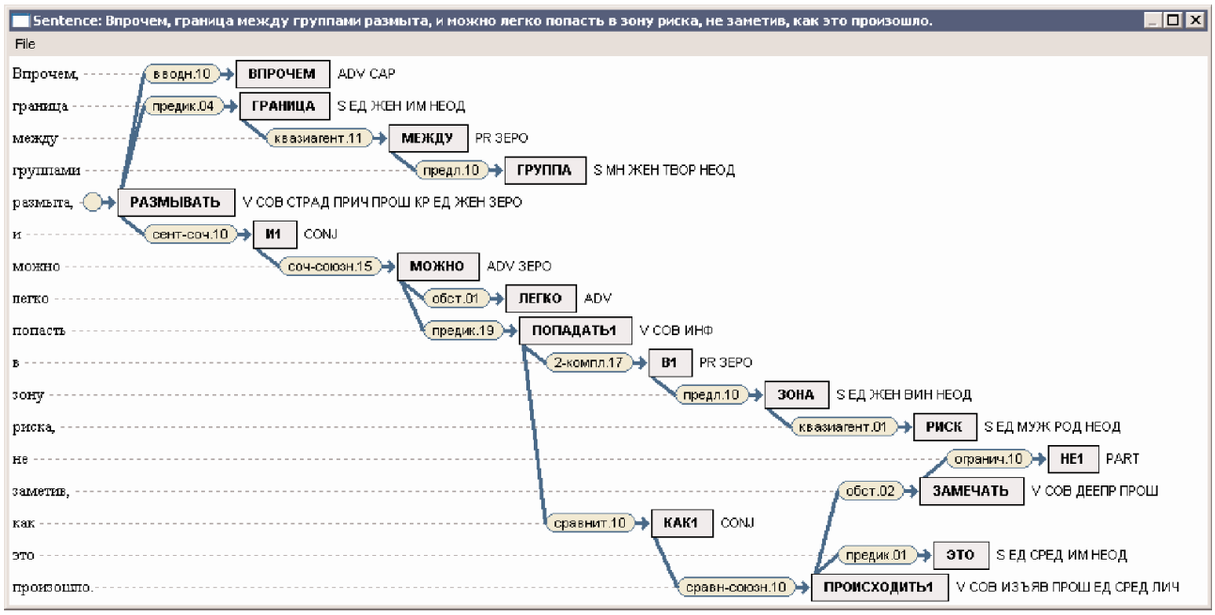


Синтаксическая структура, полученная в экспериментальном режиме ЭТАПа, выглядит так: (1б)

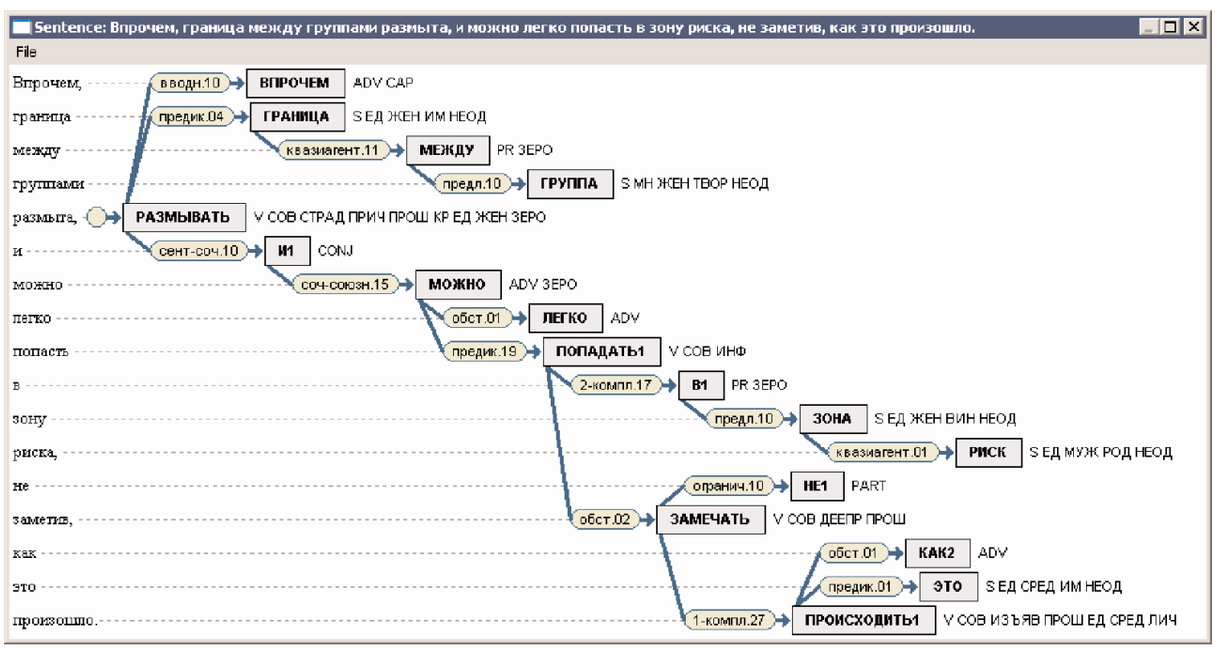
022-Drujkin_Cinman:_Layout 1 18.05.2008 14:50 Page 152

Нетрудно убедиться в том, что, хотя обе интерпретации (1) – (1а) и (1б) – с лингвистической точки зрения совершенно законны, (1б) выглядит предпочтительней. Лучший результат в эксперименте был достигнут за счет того, что “сильная” цепочка сочинительных связей сформирована для конфигурации прилагательное + союз + прилагательное, но не для конфигурации пары “прилагательное + союз + причастие, управляющее чем-либо”.
(2) *Впрочем, граница между группами размыта, и можно легко попасть в зону риска, не заметив, как это произошло.*

Основной режим ЭТАПа построил для (2) структуру (2а):



в то время как экспериментальный режим построил другую структуру: (2б)



В (2б), в отличие от (2а), заключительный фрагмент предложения (*как это происходило*), правильно подчинен глаголу *замечать* и формирует при нем придаточное дополнительное, вводимое союзным словом *как.* Это произошло благодаря тому, что синтагма-дублер *1-компл.27* породила сильную гипотезу, устанавливающую связь между глаголом *замечать* и вершиной этого придаточного *происходить.* В основном же режиме структура (2а) интерпретирует этот фрагмент как сравнительный оборот при глаголе *попадать* – законно, но неестественно.

022-Drujkin_Cinman:_Layout 1 18.05.2008 14:50 Page 153

4. Выводы

Результаты нашего ограниченного эксперимента по ранжированию синтаксических гипотез показали, на наш взгляд, перспективность этой работы. Дальнейшее направление эксперимента – увеличение количества “сильных синтагм” и отладка “сильного синтаксиса” на большом массиве. Мы рассчитываем, что относительно ограниченная модификация синтаксического анализатора приведет к заметному улучшению качества синтаксического анализа. Существенным здесь, на наш взгляд, является то обстоятельство, что экспериментатору-лингвисту – в отличие от самой современной корпусной статистики – оказываются доступны синтаксические конфигурации любой сложности и достаточно большого объема, которые он способен оценивать, опираясь на все более детальные знания о синтаксисе языка (которые в изобилии – в том числе и в форме своеобразного отрицательного языкового материала - предоставляет автоматически действующая система).

Список литературы

1. Ю.Д.Апресян, И.М.Богуславский, Л.Л.Иомдин, А.В.Лазурский, Н.В.Перцов, В.З.Санников, Л.Л.Цинман. Лингвистическое обеспечение системы ЭТАП-2. // М.: Наука, 1989, 296 с.

2. Jury D. Apresjan, Igor M. Boguslavsky, Leonid L. Iomdin, Alexander V. Lazursky, Vladimir Z. Sannikov, Victor G. Sizov, Leonid L. Tsinman. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT // MTT 2003, First International Conference on Meaning – Text Theory (June 16-18 2003). Paris: Ecole Normale Superieure, 2003. P. 279-288.

3. Ю.Д.Апресян, И.М.Богуславский, Б.Л.Иомдин, Л.Л.Иомдин, А.В.Санников, В.З.Санников, В.Г.Сизов, Л.Л.Цинман. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка 2003–2005 г. (результаты и перспективы). М: «Индрик», 2005. С. 193-214.

4. Jury D. Apresjan, Igor M. Boguslavsky, Leonid L. Tsinman. Lexical Functions in Actual NLP-Applications // Selected Lexical and Grammatical Issues in the Meaning–Text Theory. In honour of Igor Mel’čuk. (Ed. by Leo Wanner). John Benjamins, Studies in Language Companion. Series 84. 2007. P. 199-230. ISBN 978 90 272 3094 2.

023-Ermakov_A:_Layout 1 18.05.2008 14:50 Page 154

АВТОМАТИЗАЦИЯ ОНТОЛОГИЧЕСКОГО ИНЖИНИРИНГА В СИСТЕМАХ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ ТЕКСТА

AUTOMATIZATION OF AN ONTHOLOGICAL ENGINEERING FOR SYSTEMS OF KNOWLEDGE MINING IN TEXT

Ермаков А.Е*. (ermakov@rco.ru)*
ООО "ЭР СИ О", Москва

Доклад посвящен вопросам использования онтологий в системах извлечения знаний из текста. Рассматриваются особенности онтологий, используемых в таких системах. Предлагается методика автоматизированного построения онтологии, когда термины предметной области и связи между ними первоначально выделяются при помощи методов компьютерного анализа текста.

Онтологии в системах извлечения знаний

Перед системами извлечения знаний из текста сегодня встают насущные практические задачи, появление которых стимулировано развитием Интернета, содержащего огромное количество текстовой информации - реальные элементы утилитарного знания, полученные людьми в результате не только их профессиональной, но и бытовой деятельности. К таковым задачам, по мнению автора, относятся:

- Поиск и извлечение элементов знания, явно присутствующих в текстовой коллекции в виде: а) утверждения (*лекарство Антипилин – полная ерунда*; *наиболее вероятная причина свиста под капотом автомобиля в сырую погоду – слабое натяжение ремня генератора*); б) факта (*после принятия Антипилина может подниматься давление*; *летом 2006 фирма Пежо отозвала 20000 автомобилей из-за возможного возгорания в системе электроусилителя руля*).

- Порождение сложного знания путем обработки элементов знания следующими способами: а) генерация нового знания как цепочки логического вывода из элементарных утверждений и/или фактов, например: *продукт X некачественный* (утверждение), *X - продукт компании Y в 1997* (факт), *Z - технический директор компании Y с 1996 по 1998 годы* (факт), следовательно, *Z - плохой руководитель* (знание); б) эксплицирование обобщенного знания, скрытого в совокупности частных утверждений и/или фактов, например, порождение выводов типа *препарат Антипилин имеет меньше побочных эффектов, чем Глипирон* (на основании анализа отзывов больных) или *Автомобили Форд Фокус ломаются чаще, чем Мицубиси. Типичная причина поломок автомобиля Форд Фокус – засорение бензонасоса* (на основании анализа отзывов владельцев автомобилей).

Согласно определению Т. Грубера, онтология - это спецификация концептуализации предметной области [1]. Это формальное и декларативное представление, которое включает словарь понятий и соответствующих им терминов предметной области, а также логические выражения (аксиомы), которые описывают множество отношений между понятиями. Для описания отношений в онтологиях используются весь арсенал формальных моделей и языков, разработанных в области искусственного интеллекта – исчисление предикатов, системы продукций, семантические сети, фреймы и т.п. Таким образом, модный сегодня термин "онтология" оказался близок по значению к термину "искусственный интеллект", а термин "онтологический инжиниринг" явился синонимом термина "инженерия знаний". На сегодняшний день существует не менее десятка зарубежных систем, относимых к классу инструментов онтологического инжиниринга, которые поддерживают различные формализмы для описания знаний и используют различные машины вывода из этих знаний. Наиболее известные из них – это Protégé (http://protege.stanford.edu), CYC (http://www.cyc.com), KAON2 (http://kaon2.semanticweb.org), OntoEdit (http://www.ontoprise.de/products/ontoedit), KADS22 (http://hcs.science.uva.nl/projects/kads22/index.html). Хороший обзор таких систем представляет собой работа [2]. Среди уже разработанных онтологий наиболее известными и объемными являются CYC (http://www.cyc.com) и SUMO (http://www.ontologyportal.org/).

Перекочевав в смежные с искусственным интеллектом области, термин онтология стал популярен в области систем машинного анализа текста, где в большинстве случаев используется в узком значении – в качестве синонима термина "тезаурус" или "классификатор" – и представляет собой просто словарь понятий (концептов), каждому из которых соответствует синонимический ряд терминов, плюс иерархическую структуру

023-Ermakov_A:_Layout 1 18.05.2008 14:50 Page 155

взаимосвязей между ними типа “часть-целое” или “общее-частное”. Такие “онтологии в слабом смысле” используются для формулировки запросов к поисковой машине, для автоматической классификации (категоризации) текстов, и, пожалуй, на этом все. Работающих прикладных программ, относимых к классу систем извлечения знаний из текста и использующих “онтологии в сильном смысле”, т.е. методы искусственного интеллекта, способные нетривиально перерабатывать извлеченные из текста элементы знаний (интерпретировать, обобщать, выявлять зависимости, прогнозировать и т.п.), сегодня не существует, во всяком случае, для русского языка. Такое ограниченное использование онтологий обусловлено, на взгляд автора, двумя факторами. Во-первых, слабым распространением систем лингвистического анализа текста, способных интерпретировать синтаксические отношения между словами и потому действительно извлекать знания как некие нетривиальные элементы, обладающие внутренней структурой, пригодные для нетривиальной смысловой обработки искусственным мозгом – такие системы на мировом и российском рынках только начали появляться в последние несколько лет (Net Owl, Attensity, RCO Fact Extractor) и еще не успели «обрасти« приложениями. Во-вторых, относительно низкой достоверностью автоматически извлекаемых из текста утверждений и фактов, что обусловленно как несовершенством алгоритмов анализа текста, так и качеством источников информации, поскольку практически интересно извлечение знаний не из научной литературы, которая уже представляет конгломерат знания, а из текстовых “помоек”, к каковым относятся социальные сети Интернет, современные СМИ, и даже архивы научно-технических отчетов.

Другая особенность применения онтологий в системах извлечения знаний из текста – необходимость иметь дополнительную лингвистическую составляющую как для распознавания различных способов обозначения понятий (синонимичные термины), так и для семантической интерпретации разнообразных языковых конструкций в отношения между этими понятиями (синонимичные лексико-грамматические конструкции).

В итоге, для систем извлечения знаний из текста наиболее типичной является онтология “в слабом смысле” с относительно бедной концептуальной, но чрезвычайно богатой лингвистической составляющей.

Онтологический инжиниринг как объект автоматизации

Объединяя стандартные операции, выполняемые при формировании концептуальной составляющей онтологии [3,4,5], с теми операциями, которые диктуются требованиями к лингвистической составляющей, можно сформулировать перечень действий, подлежащих выполнению экспертом в ходе онтологического инжиниринга:

Формирование концептуальной схемы онтологии на основании профессиональных знаний в предметной области:

а) отбор базовых понятий-концептов. Например: *автомобиль, узел, тип кузова, год выпуска, пробег, техническое обслуживание, надежность, проходимость, экономичность.*

б) классификация базовых понятий с формированием абстрактных понятий – имен классов: типов объектов, их характеристик, ситуаций с их участием. Например: понятия - типы объекта: *автомобиль*, *узел автомобиля*; понятия - типы атрибутов объекта: *год выпуска*, *пробег*, *производитель*; понятия - типы характеристик объекта: *внешний вид*, *комфорт*, *ходовые качества*, *надежность*, *безопасность*; понятия - типы ситуаций (включая роли участников): *поломка (автомобиль, узел, причина)*, *техническое обслуживание (автомобиль, место, причина, стоимость, время ожидания)*,

в) определение возможных отношений понятий. Например: *автомобиль->{описывается}->атрибут, автомобиль->{содержит в себе}->узел, узел->{содержит в себе}->узел, объект->{характеризуется}->характеристка, характеристка->{характеризуется}->характеристка, объект->{ситуация}->объект;* и т.д.

2. Формирование фактического терминологического наполнения онтологии – соотнесение всех терминов предметной области с понятиями в концептуальной схеме, в ходе чего:

а) расширяется словарь понятий за счет наращивания онтологии «в глубину», если онтология предполагает родо-видовые связи (общее->частное, часть->целое) между понятиями одного класса, например: *узел автомобиля->двигатель->система зажигания->траблер->бегунок*, *ходовые качества->управляемость->склонность к сносу передней оси*;

б) для каждого понятия формируется словарь возможных терминов-значений: *производитель автомобиля* = {*АвтоВАЗ*, *Шевроле США*, *Шевроле Украина*, *Шевроле Корея*, …}, сила двигателя={*сильный*, *слабый*}

3. Формирование лингвистической составляющей:

а) фиксируются синонимичные обозначения каждого понятия или значения (термины): *Митсубиси = Мицубиши = Mitsubishi*, *двигатель = мотор = движок, маломощный = слабый = хилый*

023-Ermakov_A:_Layout 1 18.05.2008 14:50 Page 156

б) описываются способы выражения отношений из онтологии в языке – типовые лексико-грамматические конструкции, для чего используется соответствующий лингвистическому анализатору формализм, например [6]. Так, отношение *объект->{характеризуется}->характеристка* может выражаться в тексте из Интернета такими конструкциями: *слабый двигатель, мотор – слабак, малая мощность двигателя, движок имеет небольшую мощность, движок еле тянет, автомобиль с трудом разгоняется*, *тачка не прет*, и многими другими.

Автоматизация онтологического инжиниринга предполагают такую организацию этого процесса, при которой первоначальный перечень терминов предметной области и структура их взаимосвязей автоматически выявляются программными средствами на основании статистической обработки результатов лингвистического анализа коллекции текстов, после чего верифицируются и структурируются экспертом в соответствии с его имплицитной моделью знаний и прагматическими требованиями прикладной системы, для которой разрабатывается онтология.

С теоретической точки зрения, эффективность такой автоматизации онтологического инжиниринга обуславливается следующими факторами:

- В ходе просмотра конкорданса предметной области (частотного лексикона со взаимосвязями и контекстом) у эксперта активизируются соответствующие элементы его персональной модели знаний, что стимулирует эксплицирование и вербализацию этой модели;
- Концептуальная модель, формируемая с учетом фактического текстового материала, является актуальной, так как индивидуальная модель знаний эксперта в ходе эксплицирования автоматически верифицируется и стандартизируется в соответствии с общепринятыми представлениями;
- Легко формируется актуальное терминологическое наполнение, в том числе профессиональный сленг.

Алгоритмический арсенал для обработки текста

Технические решения, предлагаемые здесь к использованию при автоматизации формирования онтологии, основываются на следующей алгоритмической базе:

- способе генерации всех грамматически правильных словосочетаний – элементов смысла текста – на основании синтаксического анализа предложений с последующим обходом сети синтактико-семантических отношений. Соответствующие правила описаны в работе [7].

- способе установления ассоциативно-статистических связей между терминами, который основан на подсчете частоты их совместной встречаемости в рамках одной структурной единицы текста, обычно предложения. При этом в качестве вероятности наличия смысловой связи между терминами A и B можно рассматривать как абсолютную частоту их совместной встречаемости F(A,B), так и ее отношение к максимальной из полных частот встречаемости F(A) или F(B), поскольку отношение F(A,B) / F(A) есть условная вероятность появления термина A совместно с термином B.

- синтаксическом способе установления связей, который предполагает выявление терминов, связанных с другими терминами на основе определенных типов связей в предложении или даже целых лексико-синтаксических конфигураций, определяемых требуемыми шаблонами [6]. Сложность используемых лексико-синтаксических шаблонов определяется наличием априорных знаний о типовых способах языкового описания отношений в предметной области. В наиболее простом и типичном случае возможен анализ на основании самых общих синтаксических шаблонов:

- Согласованное определение (прилагательное, причастие) выражает атрибут, качество объекта: *мощный двигатель*, *стучащая подвеска*;
- Признаковое существительное, при котором объект упоминается в позиции несогласованного определения, выражает атрибут, качество объекта: *мощность двигателя*, *мягкость подвески*;
- Событийное (обычно отглагольное) существительное, при котором объект упоминается в позиции несогласованного определения, выражает ситуацию, в которой участвует объект: *работа двигателя*, *стук подвески*;
- Существительное или прилагательное, связанное с объектом глаголом-связкой или стоящее в позиции субстантивного сказуемого, выражает атрибут, качество объекта: *двигатель – (был) слабак*, *подвеска является мягкой*;
- Полнозначный глагол или событийное существительное, при котором объект выступает в роли актанта, представляет ситуацию (действие, процесс, состояние), в которой участвует объект: *двигатель тянет*, *перебирать двигатель*, *стук в подвеске*.
- Наречие при глаголе, при котором объект упоминается в позиции субъекта, косвенно выражает характеристику объекта через его действие: *двигатель тянет слабо*, *подвеска мягко покачивается*.

023-Ermakov_A:_Layout 1 18.05.2008 14:50 Page 157

Как видно, достоинством синтаксического способа является высокая точность выявления связей. Достоинством ассоциативно-статистического способа является его универсальность, которая заключается в отсутствии необходимости априорных предположений о структуре возможных синтаксических связей между терминами, и устойчивость к стилю текста, позволяющая выявить ассоциативные связи даже на грамматически некорректном тексте или тексте особого стиля, к каковым часто относятся сообщения из Интернета.

Методика автоматизированного построения онтологий

Описываемая далее методика автоматизации операций, выполняемых экспертом в ходе разработки онтологии, базируется на идее итерационного выделения из коллекции текстов вначале наиболее простых и часто упоминающихся «сущностей» предметной области, а затем все более сложных, на основании определенных критериев их связи (сочетаемости) с более простыми сущностями, зафиксированными экспертом в ходе обработки результатов предыдущих итераций.

Методика состоит из следующих шагов:

Этап 1. Построение словаря терминов – обозначений “сущностей” предметной области.

Для каждого предложения текста производится синтаксический анализ с получением дерева синтаксических зависимостей между составляющими предложения. Дерево зависимостей преобразуется в сеть синтактико-семантических отношений. На основе обхода сети синтактико-семантических отношений производится синтез термино-подобных словосочетаний [7].

Для всего корпуса текстов составляется словарь термино-подобных словосочетаний, обозначающих неодушевленные и/или одушевленные предметы – именных групп, в которых главным словом являются предметные существительные. На этом этапе в словарь не включаются глаголы, прилагательные и образованные от них существительные, которые могут представлять ситуации и свойства, связанные с объектами предметной области. Для каждого словосочетания запоминается небольшой набор ссылок на предложения текста – цитат.

Фильтрация и сортировка словаря. Для каждого термина словаря - подсчет его полной и независимой частоты встречаемости. Отношение полной и независимой частот встречаемости позволяет учесть иерархию смыслов, которая выражается в уровне синтаксической зависимости одних элементов словосочетаний от других. Например, смыслы, входящие в состав словосочетания *натяжение ремня генератора*, не равнозначны: речь идет в первую очередь о *натяжении*, затем о *ремне*, и лишь опосредованно затрагивает *генератор*. В то же время цельный смысл *натяжение ремня генератора* более информативен, чем *натяжение ремня*, а *ремень генератора* информативнее, чем *ремень*, так как включает в себя конкретизирующие элементы. В итоге, те слова и словосочетания, для которых отношение величин «частота независимой встречаемости» (не в составе других словосочетаний) и «полная частота встречаемости» оказывается близко к нулю, могут быть отброшены как неполные части устойчивых терминов.

Иерархическая группировка элементов словаря на основе лексической вложенности слов и словосочетаний. Например, два подмножества из множества словосочетаний *коробка передач*, *автоматическая коробка передач*, *механическая коробка передач*, *автоматическая коробка*, *механическая коробка*, *задняя передача*, *высокая передача*, *низкая передача* могут быть сгруппированы как по общему существительному *коробка*, так и по общему существительному *передача*, в результате чего для эксперта определяются два возможных входа в словарь: *коробка = { коробка передач*, *автоматическая коробка передач*, *механическая коробка передач*, *автоматическая коробка*, *механическая коробка }* и *передача = { коробка передач*, *автоматическая коробка передач*, *механическая коробка передач*, *задняя передача*, *высокая передача*, *низкая передача }*

Верификация/уточнение/пополнение построенного словаря терминов (обозначений объектов) экспертом в предметной области, в том числе фиксация синонимичных обозначений одних и тех же объектов.

Этап 2. Расширение словаря терминов именами ситуаций и свойств объектов предметной области.

Для каждого ранее зафиксированного термина-объекта предметной области - поиск слов (словосочетаний), связанных связями типа “объект-атрибут” и “объект-ситуация”, на основании шаблонов, задающих соответствующие конфигурации синтаксических связей.

Формирование общего словаря терминов – объектов, их атрибутов и ситуаций с их участием, группировка элементов словаря на основе взаимосвязей, выделенных на шаге 1, установление ссылок на предложения текста (цитат). Результирующий словарь представляет собой семантическую сеть взаимосвязанных сущностей трех классов, вход в которую возможен от частотного словаря имен объектов, атрибутов или ситуаций, а переход по связям между сущностями сопровождается возможностью просмотра текста, в котором связь раскрывается.

Исследование семантической сети экспертом в предметной области и окончательное формирование концептуальной составляющей онтологии (этап 1 процесса онтологического инжиниринга) - определение

023-Ermakov_A:_Layout 1 18.05.2008 14:50 Page 158

абстрактных понятий (классов объектов, их свойств и ситуаций) с определением типизированных отношений между сущностями этих классов; окончательное формирование фактического наполнения онтологии (этап 2 процесса онтологического инжиниринга) - соотнесение всех терминов словаря с понятиями в схеме онтологии, в том числе фиксация синонимичных обозначений свойств и ситуаций, определение возможных иерархических отношений между сущностями одного класса.

Если словарь объектов еще не был полностью сформирован (на этапе 1 экспертом была проанализирована только высокочастотная часть словаря), то возможно повторение шагов 1-3 с выделением новых упоминающихся объектов, связанных с уже известными свойствами и ситуациями из предметной области.

Этап 3. Описание способов выражения отношений из онтологии в языке – типовых лексико-грамматических конструкций.

Выявление множества ассоциативно-статистических связей между всеми терминами предметной области, для которых существует связь в онтологии. Ассоциативно-статистическая связь устанавливается между терминами, совместно упоминавшимися в предложениях текста не менее заданного числа раз.

Построение списков цитат из текста для каждого типа связей из онтологии, с предварительным отсевом статистически малодостоверных связей и тех связей, которые выражаются уже известными способами и могут быть выделены на основании синтаксических шаблонов (Шаг 1 Этапа 2).

Исследование списков цитат экспертом для фиксации новых способов выражения в тексте отношений из онтологии – новых лексико-грамматических конструкций, используемых впоследствии для настройки лингвистического обеспечения системы автоматического извлечения знаний из текста.

Заключение

Экспериментальная проработка и успешная апробация методики проводилась специалистами компании “ЭР СИ О” в ходе построения онтологии для предметной области “Автомобили”. Онтология предназначена для оценки конкретных марок автомобилей с точки зрения характеристик (*положительная/отрицательная*) их потребительских свойств, высказываемых в отзывах потребителей, размещенных в Интернете. При составлении онтологии использовался реальный языковой материал, полученный из автомобильных сообществ блога “Живой Журнал” (http://www.livejournal.ru/auto) – около 30 Мбайт текстовых сообщений. Результирующая онтология содержит более 1200 терминов (не считая конкретных марок автомобилей), из которых 211 представляют собой наименования узлов автомобиля (*движок*, *коробка передач, ходовая часть*); 71 - наименование их свойств (*ходовые качества*, *комфорт*, *надежность*, *стоимость содержания*); 882 - возможные наименования оценок характеристик узлов и свойств, включающие прилагательные, существительные, глаголы и наречия (*крутой*, *поломка*, *глючить*, *отстойно*), 37 эмоциональных характеристик (*любить*, *жалоба*, *плеваться*). Возможные связи в предложении между классами терминов из онтологии описываются 150 лексико-грамматическими шаблонами. В результате для каждой модели автомобиля в блогах удается «выловить» положительные и отрицательные отзывы, классифицировав их по темам «за что хвалят/ругают».

Список литературы

1. Gruber T. R. A translation approach to portable ontologies // Knowledge Acquisition, 1993, V. 5(2), P.199-220.

2. Овдей О.М., Проскудина Г.Ю. Обзор инструментов инженерии онтологий // Электронные библиотеки – Москва: Институт развития информационного общества, т.7 вып.4, 2004. – Электронный журнал, посвященный созданию и использованию электронных библиотек. – (http://www.elbib/).

3. Гладун А., Рогушина Ю. Онтологии в корпоративных системах // Корпоративные системы, – 2006. – № 1. – с. 41-47.

4. Гаврилова Т.А. Использование онтологий в системах управления знаниями // Труды международного конгресса «Искусственный интеллект в XXI веке», Дивноморское, Россия, М., Физматлит. 2001 - с. 21-33.

5. Гаврилова Т.А. Извлечение знаний: лингвистический аспект // Корпоративные системы (Enterprise Partner), 2001. - № 10 (25). - с. 24-285.

6. Киселев С.Л., Ермаков А.Е., Плешко В.В. Поиск фактов в тексте естественного языка на основе сетевых описаний // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’2004. – Москва, Наука, 2004. – С. 282-285.

7. Ермаков А.Е. Эксплицирование элементов смысла текста средствами синтаксического анализа-синтеза. // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’2003. – Москва, Наука, 2003. - С. 136-140.

024-Zalizniak_A:_Layout 1 18.05.2008 14:51 Page 159

ТЕКСТЫ А.ПЛАТОНОВА КАК ЛИНГВИСТИЧЕСКИЙ ИСТОЧНИК
A.PLATONOV'S TEXTS AS A LINGUISTIC SOURCE

Зализняк Анна А. *(anna-zalizniak@mtu-net.ru)*
Институт языкознания РАН

В статье демонстрируется, как языковые аномалии в текстах А.Платонова, которые до сих пор исследовались исключительно как источник сведений о поэтическом мире автора, могут быть использованы в качестве источника нетривиальной информации о семантических, сочетаемостных и категориальных свойствах слов русского языка.

Язык Андрея Платонова представляет собой удивительный феномен. Одна из его выдающихся особенностей состоит в том, что этот язык одновременно легко узнаваем и очень трудно имитируем. Действительно, почти любая фраза из произведений Платонова (я имею в виду художественную прозу 20х – 30-х годов) может быть опознана как принадлежащая этому автору. Однако попытки имитации особенностей его стиля обычно оказываются неудачными. Так, например, сконструированные Т. Б. Радбилем фразы *Дай мне прозрачную жидкость, предназначенную для питья*; *Человек сел на среднюю часть спины лошади* в качестве образцов языка Платонова [Радбиль 2007: 281], на мой взгляд, таковыми никак служить не могут: именно Платонов так никогда бы не сказал. Заметим, что в нашумевшем в свое время сборнике пародий «Парнас дыбом» [Паперная, Розенберг, Финкель 1927], задуманном как своего рода исследование стиля разных писателей (от Гомера до Окуджавы), пародии на Платонова нет – хотя особенность его языка настолько бросается в глаза, что, казалось бы, идея пародии напрашивается сама собой[1].

Другой не менее удивительной особенностью языка Платонова является его понятность (ср. [Дмитровская 1988: 108], [Стернин 1999: 154]). Более того, он ловит в ловушку читателя, у которого создается впечатление, что для того, чтобы говорить на языке и быть понятым, вовсе не нужно соблюдать правила. Это, конечно же, ложное впечатление. Как мне кажется, именно по этому ложному следу направляет Платонов тех исследователей, которые трактуют его язык как «язык мысли», «язык смысла», праязык, дологический, внелогический, мифологический и т.д. язык, не различающий никаких категорий (которые различает обычный язык), не имеющий никаких презумпций и т.д. Это ложное впечатление – хотя, возможно, именно оно и входит в intentio auctoris (я имею в виду трихотомию Умберто Эко «intentio auctoris – intentio operis – intentio lectoris» [Eco 1990], к которой я еще вернусь).

Впечатление от текстов А. Платонова очень точно, на мой взгляд, описал Ю.И.Левин в статье [Левин 1990], обозначив его как «единство формы и содержания». Действительно, вышеупомянутое единство в текстах Платонова, столь велико, что занимаясь *синтаксисом*, исследователь, сам того не замечая, переходит к *смыслу* и – *далее*. Именно этому «далее» (т.е. онтологии и мифологии его поэтического мира), и посвящена фактически вся литература о Платонове, – которая на сегодня уже практически необозрима.

Цель настоящей работы иная. Она состоит в обнаружении и описании тех правил русского языка, которые язык Платонова нарушает (ср. [Падучева 2007]). Дело в том, что эти правила иногда оказываются столь тонкими и сложно формулируемыми, что без помощи Платонова мы, возможно, никогда бы и не догадались о их существовании. Приведу пока только два примера. Первый: *мужик ложился вниз и* ***как можно скорее плакал*** (К)[2]. Здесь причина аномальности состоит в том, что глагол *плакать* не является видовым коррелятом к *заплакать*, потому что не может заменять глагол *заплакать* в контексте обязательной имперфективации (см. [Зализняк, Шмелев 2001]), в данном случае – итеративном, ср. допустимое, например, *...ложился вниз и как*

[1] Однако «чисто платоновские» выражения встречаются у других авторов, порой весьма неожиданным образом. Приведу два примера:

> Лида кричала, что ему, старику, следует думать о своей науке, а не о паршивых девицах,... что она поражена *низменными уклонами* своего отца (М.Зощенко. Возвращенная молодость)

Здесь ненормативное словосочетание *низменные уклоны* возникло из «наложения» стандартного *низменные наклонности* и актуального в ту эпоху слова *уклон* (в политическом смысле). Второй пример – фраза *Вы в работе должны показывать весь ум своего мышления* из пародированной речи советского начальника в повести Владимира Кантора «Крокодил».

[2] При отсылке к произведениям Платонова используются следующие сокращения: К – «Котлован», Ч – «Чевенгур», СИ – «Семья Иванова».

024-Zalizniak_A:_Layout 1 18.05.2008 14:51 Page 160

можно скорее ***засыпал***. Если бы глагол *плакать* был видовым коррелятом для *заплакать*, то он обладал бы, в частности, и способностью этого глагола приобретать необходимую в данном контексте сему контролируемости – как происходит, например, в случае с глаголом *засыпать*, который, будучи коррелятом к *заснуть*, с той же легкостью приобретает под влиянием контекста дополнительную сему сознательного приложения усилий. При этом следует добавить, что видовая коррелятивность глагола *заплакать* составляет определенную проблему для русской аспектологии, и цитированный выше пример из «Котлована» является важным аргументом в пользу непарности этого глагола.

Второй пример: *тщетно и смутно было вокруг.* Не говоря о том, что значение этого выражения абсолютно понятно несмотря на его абсолютную аномальность, здесь обращает на себя внимание следующее обстоятельство, касающееся слова *вокруг*. Что *вокруг* не может быть *тщетно* и *смутно*, не вызывает ни сомнения, ни удивления (*смутно* бывает только *на душе*, и то это несколько сдвинутое употребление, а *тщетно* даже и на душе, строго говоря, быть не может). Однако при ближайшем рассмотрении оказывается, что *вокруг* не может быть также, например, *холодно* или *жарко*. Можно сказать: *здесь тоскливо*; *здесь скучно*, но нельзя *??Вокруг было тоскливо*, **Вокруг было скучно*. Т.е. слово *вокруг* привносит дополнительную составляющую физического пространства, которой нет в *здесь*. Таким образом, оказывается, что *вокруг* не может быть никак в смысле внутреннего состояния, а может быть только то, что человек видит (или слышит) в окружающем его пространстве, ср. *вокруг было пусто, темно, светло, полно народу, безлюдно, тихо и красиво, тихо и спокойно, шумно* и т.п.[3] Здесь напрашивается вывод о том, что внешнее по отношению к человеку пространство оказывается, в мире Платонова, внутри человека, но он уже сделан исследователями; можно считать, что употребление слова *вокруг* его подтверждает, но меня в данном случае интересует значение этого слова в русском языке: то, что слово *вокруг* обозначает лишь воспринимаемое человеком зрительно пространство – это неожиданный, на мой взгляд, результат.

Итак, в данной работе ставится нетрадиционная для платоноведения задача: не «в мир Платонова через его язык» (ср. название книги [Михеев 2003], а также статьи [Левин 1990] и книги [Dhooge 2007]), а в некотором смысле наоборот, «в мир русского языка через язык Платонова». Другими словами, я хочу обратить внимание лингвистов на то, что язык Платонова – это бесценный источник сведений об устройстве русского языка.

Все, о чем я буду говорить, не выходит за пределы intentio operis. Предлагаемая мною модель исходит из того, что язык Платонова – это не «недо-язык» ни в каком из перечисленных выше смыслов, а что это артефакт, Kunststück, созданный (насколько намеренно, осознанно и т.д. – не имеет никакого значения, как я уже сказала, проблематика intentio auctoris вообще выводится за рамки рассмотрения) при помощи множества нарушений языковых правил и конвенций разного типа, т.е. он является вторичным, производным по отношению к стандартному русскому языку. Моя задача – лишь описать эти правила и конвенции.

Среди работ, посвященных собственно языку Платонова, назову три статьи, в которых производится убедительный собственно лингвистический анализ – [Левин 1990], [Кобозева, Лауфер 1990], [Бобрик 1995], книгу [Дооге 2007] и особенно диссертацию [Tsvetkov 1983], которая содержит, по-видимому, наиболее проницательный анализ механизмов Платоновских языковых нарушений, который до сих пор предлагался[4]. Вообще надо сказать (и это обстоятельство послужило одним из стимулов написания данной статьи), что чтение литературы о Платонове, поражает, с одной стороны, тем, насколько произвольно толкуются его «нестандартные словосочетания» и, с другой, насколько предлагаемые толкования иногда оказываются менее понятны, чем само толкуемое выражение. Первое обстоятельство легко объясняется в рамках рецептивной поэтики (согласно которой любое художественное произведение является «открытым» – в том смысле, что читатель имеет не меньше прав на интерпретацию текста, чем создавший его автор, ср. intentio lectoris). Что касается второго, то, на мой взгляд, в какой-то степени оно объясняется известным принципом о том, что «сложное понятней им» – который, переведя на язык науки, можно переформулировать примерно таким образом, что имплицитность некоторой части обрабатываемой информации есть ингерентное свойство того вида человеческой когнитивной деятельности, которая осуществляется при помощи языка и результаты которой до какой-то степени в нем отражены. «Вытаскивая наружу» смыслы, присутствующие в выражении в неявном виде, мы неизбежно искажаем тот объект, который мы исследуем. Другими словами, некоторые вещи в принципе нельзя выразить явно: их можно только «имплицировать» [5].

В работе [Зализняк, Левонтина 1996] уже высказывалась мысль о том, что степень имплицитности

[3] Я здесь обобщаю данные Национального корпуса русского языка.

[4] Главный недостаток работы А.П.Цветкова – ее недоступность: она существует как диссертация, защищенная в Мичиганском университете в 1983 г. (т.е. даже не в электронной версии), и больше никак. Я благодарна Бену Дооге, предоставившему в мое распоряжение сделанную им лично ксерокопию этого замечательного труда.]).

[5] Поэтому толкование довольно часто оказывается менее понятно, чем толкуемое выражение, и тому есть достаточно серьезные причины (см. обсуждение этой проблемы в [Зализняк, 2006].

024-Zalizniak_A:_Layout 1 18.05.2008 14:51 Page 161

выражения некоторого смысла в семантической структуре слова является неотъемлемой частью его значения, и в силу этого и переводной эквивалент, в котором степень имплицитности тех же смысловых компонентов иная, и перифраза на каком-либо языке (в том числе семантическом, т.е. толкование), эксплицирующая как все содержащиеся в слове семантические компоненты, так и их коммуникативный статус ipso facto искажают значение толкуемого слова. Высказанный пессимизм относительно возможностей аппарата толкований никак не означает, естественно, что от него надо отказаться (просто эту ограниченность следует осознавать).

Впрочем, целью настоящей работы и не является толкование. Цель, как уже говорилось, состоит в том, чтобы предложить научному сообществу взгляд на является язык Платонова как на источник сведений о русском языке.

Начну с очень существенной оговорки, касающейся выражения *язык Платонова*. Как мне кажется, здесь имеется одно недоразумение, связанное с тем, что слово *язык* в основном значении (*русский*, *английский* и т.д. *язык*) и в значении, реализующемся в конструкции *язык Пушкина, Шекспира, Платонова* и т.д. обозначает объекты различного семиотического типа. Это обстоятельство имеет ряд важных следствий; например, только от слова *язык* в первом значении образуется прилагательное *языковой*, и только по отношения к нему осмысленно сочетание *языковая концептуализация*. Говоря о языковой концептуализации, обычно имеют в виду то обстоятельство, что в словах и конструкциях естественного языка отражен определенный способ видения мира, отдельных его фрагментов и свойств (присущий носителям этого языка). Субъектом этой концептуализации является, в некотором смысле, сам язык – а именно, в том смысле, что естественный язык существует «сам по себе» и функционирует не контролируемым человеком образом. С другой стороны, язык является «субъектом» в том смысле, что он «навязывает» заложенную в нем концептуализацию мира говорящим на этом языке (которые, сами того не замечая, принимают некоторые представления о мире как сами собой разумеющиеся лишь в силу того, что они пользуются данным языком).

Объект, обозначаемый как *язык Платонова,* совсем другого рода. Прежде всего, это артефакт. И специфический способ концептуализации мира, из него нами извлекаемый, принадлежит не языку, а человеку, его создавшему. Тем самым это совокупность смыслов определенного типа, созданная одним человеком определенным способом; это можно назвать «концептуализацией Платонова»; его «поэтическим миром» или просто «миром», но это не есть «языковая концептуализация».

Мы не будем останавливаться на классификации типов аномалий у Платонова (см. об этом, в частности, [Радбиль 2006], [Dhooge 2007]) и ограничимся лишь сочетаемостными аномалиями в широком смысле. Специально обратим внимание на то, что у Платонова практически нет неологизмов – что отличает его от большинства других языковых экспериментаторов той эпохи.

В языке Платонова имеется несколько механизмов, порождающих сочетаемостные аномалии (некоторые из них могут действовать одновременно): (1) **подстановка** (вместо некоторого слова его синонима), и это основной механизм, определяющий облик платоновского языка; далее: (2) **склеивание двух конструкций** в одну; (3) **анаколуф** (или **силлепсис**: когда одно слово употреблено сразу в двух значениях, потому что оно связано с двумя словами, которые требуют двух разных значений, типа *шел дождь и два студента*, но гораздо тоньше; иногда это вообще почти незаметно, как в *провожали с любовью, сожалением, с песнями и вином* (СИ)); (4) **рассогласование категориальных признаков** слов синтаксически связанных слов; (5) **ресемантизация идиом** (ср. *он стал мне близким* (СИ) вместо стандартного *мы стали близки*).

Механизм **подстановки** состоит в следующем. Некоторые два слова могут быть в каком-то типе контекстов (квази)синонимичны и до какой-то степени взаимозаменимы; ср. *разлука* и *расставание*: *перед разлукой они обнялись* и *перед расставанием они обнялись*. Однако в других контекстах в стандартном русском языке употребляется только одно из двух этих слов, так, можно сказать *три года они жили в разлуке*, но нельзя **три года они жили в расставании*. А Платонов именно так и говорит. Иными словами, на основании того, что некоторое слово может быть заменено на свой синоним в некотором одном контексте, оно заменяется на этот синоним в другом, в котором оно в стандартном русском языке употреблено быть не может (при этом допускается еще и синтаксическая деривация по Куриловичу и некоторые другие типы трансформаций). В этом и состоит прием подстановки[6]. Так, словосочетание *разлука с жизнью* возникает из стандартного сочетания *расстаться с жизнью* в результате применения двух операций: 1) синтакисческой деривации (*расстаться* → *расставание*) и 2) замены одного синонима другим (*разлука* вместо *расставание*).

Другой пример: *давно не находился в реке* (Ч). Сочетание *быть в* синонимично *находиться в* определенном круге контекстов, ср. *Петя сейчас в Германии = Петя сейчас находится в Германии*; *я никогда не был в столь дурацком положении = я никогда не находился в столь дурацком положении*. Однако, например, в предложении *Петя не был в Англии* заменить *был* на *находился* никак нельзя (а Платонов именно это и делает).

[6] Этот прием был впервые выявлен и описан А.П.Цветковым [Tsvetkov 1983]. Ср. также операцию «замены» в [Кобозева, Лауфер 1990].

024-Zaliznjak_A:_Layout 1 18.05.2008 14:51 Page 162

Все это наводит на размышления об особенностях семантики и порождаемых ею ограничениях сочетаемости глагола *находиться*: утверждения о том, что предикат находиться «абстрактный» и требует «абстрактной локализации» явно недостаточно, здесь необходим гораздо более тщательный анализ.

Прием **склеивания** конструкций очень близок к подстановке; собственно, он отличается от нее лишь тем, что замещенный квазисиноним сохраняет свой «след» в виде модели управления, несвойственной реально присутствующему в тексте слову-заместителю. Например: *стал заботиться над разведением костра* (К). Здесь «склеены» две конструкции: *трудиться над чем-то* и *заботиться о чем-то* (опущенное *трудиться* оставило по себе «след» в форме управления *над чем-то*)[7].

Другой пример: *ему никто не возражал здесь находиться* (К). Глагол *возражать* имеет модель управления *<против чего-то>* или *<чтобы кто-то что-то сделал>*. Замещенный глагол, управляющий инфинитивом, мог быть, например, *запрещать <кому-то что-то делать>*; здесь однако может быть «восстановлен» и какой-то другой глагол, управляющий сочетанием <дат. + инф.> (например, *препятствовать*, *не давать*); строго говоря, в данном случае (и таких много) у нас нет оснований восстановить именно тот, а не другой глагол; т.е. здесь достаточно некоторой абстракции «семантика + модель управления» в духе книги [Апресян, 1967].

Интересен следующий пример, где видна «техника» процедуры «склеивания», так как он представляет собой некий промежуточный результат: *чтобы не вспоминать и не мучиться о матери* (Ч). «Окончательный» результат был бы: *чтобы не мучиться о матери*, где слово *мучиться* возьмет себе валентности вытесненного им *вспоминать* (и именно таких примеров у Платонова очень много, ср.: *я вдалеке томился о них* (К), т.е. *думал о них* + *томился*, ср. также выше).

Частный случай склеивания представляют собой примеры типа *горевал головой в ладони* (где «склеиваемые» элементы – не квазисинонимы, а метонимически связанные концепты; здесь также одно из двух слов опускается, оставляя по себе «след» в виде зависимого компонента *в ладони*)[8]. На факт наличия «склеивания» указывает именно нарушение сочетаемости, которое может быть как совершенно очевидным, так и весьма тонким, ср. аномальность не только *у него уже загодя болел живот* (К), где речь идет о неконтролируемой ситуации, но и *присутствовали на собрании уже загодя*: *загодя*, как и *заранее*, можно лишь совершить действие, т.е. например, прийти, но не находиться в помещении.

Во всех таких случаях помимо значения двух участвующих в игре языковых единиц, из их взаимодействия обычно возникает еще некий новый, третий смысл, иногда трудноопределимый и трудноуловимый. Данное явление, само по себе весьма интересное, находится, однако, за пределами основного сюжета данной работы, поэтому мы на нем здесь останавливаться не будем.

Наибольший интерес для исследователя русского языка представляют случаи анаколуфа (которые поставляют материал для уточнения существующего разграничения лексических значений) и случаи рассогласования категориальных признаков, которые выявляют ограничения сочетаемости, иногда весьма нетривиальные.

Приведем некоторые примеры рассогласования категориальных свойств элементов словосочетания. Так, во фразе *пошел в этот город жить* (К) оно состоит в том, что *жить* – абстрактный глагол, а *идти* в конструкции с инфинитивом – конкретный (в *пошел в солдаты* – он абстрактный). Поэтому можно сказать *пошел ночевать к приятелю* (оба действия конкретные), можно *переехал жить в этот город* (оба – абстрактные), а пошел в этот город жить – нельзя.

Другой пример: *дошел до конца города*. Категориальное нарушение здесь состоит в том, что у *города* нет *конца*: В рамках той концептуализации мира, которая закреплена в русском языке, деревня – такой объект («длинный»), у которого есть конец (поэтому можно *дойти до конца деревни*), а город – это «круглый» объект, у которого конца нет (ср. [Рахилина, 2000: 247][9]. С другой стороны, по-русски говорят на другом конце города, *в разных концах города, изъездить город из конца в конец* и т.п. эти сочетания поддерживают обеспечивают внешнюю легитимность выражения *дошел до конца города* (но все же не снимают его аномальности).

Результатом данной работы должен стать фрагмент словаря русского языка с нетривиальными сведениями о их синтаксисе и семантике: таксономическая категория (собственная и требование на таксономическую

[7] Ср. также примеры «совмещения синонимов» из [Кобозева, Лауфер, 1990: 131]: думается разное в голову (совмещены думать и приходить в голову).

[8] Пример из статьи [Кобозева, Лауфер, 1990: 129], где такие случаи называются «двойной категоризацией»; ср. другие примеры: *пристально интересуясь рыцарем* (т.е. «пристально глядела, интересуясь»); *истребил ее* [*записку*] *на четыре части* (т.е. «истребил, разорвав»).

[9] Обратим внимание на то, что по-русски *конец города* сказать нельзя, хотя, существует дорожный знак (название города, перечеркнутое красным), обозначающий как будто именно «конец города такого-то».

024-Zalizniak_A:_Layout 1 18.05.2008 14:51 Page 163

категорию актанта; возможно, обнаружатся некоторые новые таксономические категории), некоторые нетривиальные сведения о модели управления, сведения о сочетаемости (например, что *замертво* можно только *упасть*, и так сказать можно только о живом существе), которые следовало бы включить в некий гипотетический суперсловарь русского языка, где отражены все релевантные лексические и грамматические свойства русских слов. Очевидно, что это будут необычайно ценные сведения о русском языке, которые должны быть использованы как минимум, в толковых словарях и в лингвистических базах данных (включая семантическую разметку корпусов, ср. [Кустова и др. 2005]).

Список литературы

1. Апресян Ю.Д. Экспериментальное исследование семантики русского глагола. М.: Наука, 1967.

2. Бобрик М. Заметки о языке Андрея Платонова // Wiener Slawistischer Almanach 35(1995).

3. Дмитровская М.А. «Переживание жизни»: о некоторых особенностях языка А. Платонова // Логический анализ языка. Противоречивость и аномальность текста. М., 1990.

4. Зализняк Анна А., Левонтина И.Б. Отражение национального характера в лексике русского языка (размышления по поводу книги: Anna Wierzbicka. Semantics, Culture, and Cognition. Universal Human Concepts in Culture-Specific Configurations. N.Y., Oxford, Oxford Univ. Press, 1992). // Russian Linguistics, vol. 20, 1996.

5. Зализняк Анна А., Шмелев А.Д. Типы видовой связи // Труды международного семинара Диалог-2001 по компьютерной лингвистике и ее приложениям. Т. 1. Аксаково.

6. Кобозева И.М., Лауфер Н.И. Языковые аномалии в прозе А. Платонова через призму процесса вербализации // Логический анализ языка. Противоречивость и аномальность текста. М., 1990.

7. Кустова Г.И., Ляшевская О.Н., Падучева Е.В., Рахилина Е.В. Семантическая разметка лексики в Национальном корпусе русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003-2005. М.: Индрик, 2005.

8. Левин Ю.И. От синтаксиса к смыслу и далее («Котлован» А. Платонова) // Семиотика и информатика. Вып. 30. М., 1990.

9. Меерсон О. «Свободная вещь». Поэтика неостранения у Андрея Платонова. Berkeley Slavic Specialities, 1997.

10. Михеев М.Ю. В мир А.Платонова через его язык. М., 2003.

11. Падучева Е.В. [рец.] Т.Б. Радбиль. Языковые аномалии в художественном тексте: Андрей Платонов и другие. М.: МПГУ, 2006. // Русский язык в научном освещении, №14, 2007.

12. Паперная Э.С., Розенберг А.Г., Финкель А.М. Парнас дыбом. Харьков: Космос, 1927.

13. Радбиль Т.Б. «Парадоксы неконвенциональности» и язык Андрея Платонова // Die Welt der Slaven LII, 2007, 281-298.

14. Рахилина Е.В. «Без конца и без края» // Исследования по семантике предлогов. М., 2000.

15. Стернин И.А. 1999 – «Язык смысла» Андрея Платонова // Филологические записки: Вестник литературоведения и языкознания. Вып. 13. Воронеж, 1999. С.154-161.

16. Dhooge B. Творческое преобразование языка и концептуализация мира у А.П.Платонова. Диссертация на соискание степени доктора восточно-европейских ячзыков и культур. Gent, 2007.

17. Eco U. I limiti dell'interpretazione. Milano: Bompiani, 1990.

18. Tsvetkov A.P. 1983 – The language of A.Platonov. A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Slavic Languages and Literatures) in the University of Michigan, 1983.

025-Zatsman:_Layout 1 18.05.2008 15:11 Page 164

ТЕРМИНЫ ДЛЯ ОПИСАНИЯ ПРОЦЕССОВ ПРЕДСТАВЛЕНИЯ НАУЧНО-ТЕХНИЧЕСКИХ ЗНАНИЙ В ЦИФРОВОЙ СРЕДЕ

TERMS FOR SCIENTIFIC AND TECHNICAL KNOWLEDGE REPRESENTATION IN DIGITAL SPHERE

Зацман И.М. *(im@a170.ipi.ac.ru),* ***Курчавова О.А.*** *(koa@a170.ipi.ac.ru)*
Институт проблем информатики РАН

Документы 7-й Рамочной программы Европейского Союза, принятой на период 2007-2013гг., содержат формулировки новых задач, относящихся к проблеме представления знаний в цифровой среде. В докладе анализируются ключевые положения этих формулировок. Результаты анализа используются для определения ряда терминов, предлагаемых для описания технологий кодирования научно-технических документов, процессов представления и сохранения знаний в институциональных электронных библиотеках.

Введение

Основная цель доклада заключается в определении ряда терминов, необходимых для описания процессов представления знаний в электронных библиотеках, являющихся ключевыми компонентами институциональных систем (далее по тексту - институциональные электронные библиотеки или ИЭБ). Например, для такой институциональной системы как патентная сфера электронные библиотеки являются ключевыми компонентами, так как в соответствии с действующими нормативными актами любому решению о принятии или отклонении патентной заявки должен предшествовать поиск по тематике этой заявки.

Процессы представления и сохранения знаний в институциональных электронных библиотеках имеют три следующие отличительные особенности. Во-первых, в ИЭБ период хранения документов может многократно превышать цикл жизни одного поколения аппаратно-программных средств их создания и поддержки. Поэтому, в докладе будут рассматриваться ИЭБ долговременного применения. Во-вторых, во время длительного хранения документов могут существенно измениться институциональные классификационные системы, тезаурусы, онтологии ИЭБ и регламенты их ведения. Во-третьих, в процессах представления и сохранения знаний в ИЭБ нередко необходимо различать *личностные, локальные и конвенциональные концепты, стабильные (lasting concepts) и нестабильные концепты (volatile concepts)*.

Необходимость учета перечисленных особенностей приводит к принципиально новым требованиям к системам представления знаний в ИЭБ, в том числе, к реализации технологий рубрицирования и кодирования документов ИЭБ. Для описания содержания этих требований необходимы новые понятия и термины. Примером могут служить выделенные курсивом словосочетания, которые в докладе определяются как новые элементы системы терминов, ранее определенной в работе [1]. С помощью этой терминосистемы в будущем планируется описать основные этапы процесса образования конвенциональных концептов на основе личностных и локальных концептов, представленных в классификационных системах, тезаурусах и онтологиях ИЭБ, а также описать различие в методах кодирования стабильных и нестабильных концептов.

Изменение регламента рубрицирования патентных документов

Перечисленные особенности процессов представления и сохранения знаний уже сейчас являются актуальными для разработки нового поколения технологий рубрицирования и кодирования патентных документов. С 2006 года в патентной сфере вступила в силу Восьмая редакция Международной патентной классификации (МПК). Для применения Восьмой редакции МПК (далее по тексту - МПК8) как основного средства представления и сохранения научно-технических знаний в цифровой среде[1], в том числе, в патентных электронных библиотеках, были внесены существенные изменения и в структуру МПК, и в регламент ее ведения [2].

В отличие от предыдущих семи версий, МПК8 была разделена на рубрики базового и расширенного уров-

[1] Цифровая среда – сочетание элементов цифровой вычислительной техники, средств телекоммуникации, информационно-компьютерных систем, иных цифровых средств ввода, хранения, поиска, передачи и других процессов обработки данных.

025-Zatsman:_Layout 1 18.05.2008 15:11 Page 165

ней. Рубрики базового уровня, выражающие стабильные концепты, могут пересматриваться не чаще, чем раз в три года, а рубрики расширенного уровня могут пересматриваться значительно чаще, начиная с января 2007 года. В соответствии с новым порядком вновь вводимые или изменяемые рубрики расширенного уровня подготавливаются Специальным подкомитетом Всемирной организации интеллектуальной собственности (ВОИС) по пересмотру расширенного уровня МПК с регулярностью один раз в квартал и передаются в Международное бюро ВОИС. Оно доводит их до сведения национальных и региональных патентных ведомств за три месяца до начала их ввода в действие с тем, чтобы патентные ведомства могли принять обеспечительные меры (перевод новых рубрик и подготовка их к публикации, ознакомление экспертов, начало их простановки на публикуемых документах и т.п.). Важно отметить, что для повышения эффективности поиска, наряду с обязательным рубрицированием изобретений в целом, МПК8 может использоваться и для кодирования отдельных содержательных аспектов описаний изобретений [3].

Рубрики расширенного уровня на этапе их рассмотрения Специальным подкомитетом ВОИС и согласования их содержания выражают личностные и/или локальные концепты участников обсуждения. После принятия положительного решения, ознакомления экспертов с вновь вводимыми или измененными рубриками и начала их использования в процессе рубрицирования и кодирования документов, они приобретают конвенциональный характер в пределах патентной сферы как институциональной системы.

При простановке рубрик расширенного уровня на публикуемых документах одновременно указывается дата (год и месяц), когда эта рубрика была введена в действие. По мере необходимости рубрики расширенного уровня, выражающие нестабильные концепты, могут пересматриваться чаще, чем раз в три года, пока им не будет присвоен статус рубрик базового уровня.

Патентная сфера не является единственной институциональной системой, в которой в процессах представления и сохранения знаний в цифровой среде необходимо различать личностные, локальные и конвенциональные концепты, стабильные и нестабильные концепты. Другим примером могут служить аналогичные процессы в системах информационного мониторинга в сфере науки, в которых экспертные знания об индикаторах результативности включают личностные, локальные и конвенциональные концепты [4, 5]. Еще одним примером являются электронные геобиблиотеки, вербально-образные тезаурусы которых разрабатываются на основе локальных классификационных систем геообъектов и явлений [1, 6].

Приведенные примеры иллюстрируют актуальность задачи построения системы терминов для описания процессов представления и сохранения знаний, в рамках которой можно было бы определить смысл словосочетаний «локальные (личностные) концепты», «стабильные концепты» и «нестабильные концепты». В докладе предпринята попытка предложить для перечисленных словосочетаний дефиниции на основе системы терминов из работы [1].

Проблема представления знаний в документах 7-й Рамочной программы ЕС

Важно отметить, что предлагаемые дефиниции имеют самостоятельную ценность и вне описания процессов рубрицирования и кодирования научно-технических документов, так как они необходимы для более детального описания и согласованного понимания проблемы представления знаний в цифровой среде, в том числе, понимания того варианта описания этой проблемы, который представлен в документах 7-й Рамочной программы Европейского Союза (ЕС), принятой на период 2007-2013гг.

В этих документах сформулировано восемь приоритетных направлений исследований и разработок в области информационно-коммуникационных технологий (ИКТ), включая в качестве отдельных направлений «Электронные библиотеки и их содержание» и «Перспективные ИКТ» [7, 8]. Именно в этих двух направлениях нашла отражение проблема представления и сохранения знаний в цифровой среде. При этом, в описаниях приоритетных направлений исследований и разработок, как правило, не определены ключевые термины.

В качестве примера рассмотрим несколько положений, взятых из формулировок приоритетных направлений исследований и разработок в области ИКТ. Долгосрочные цели проектов по приоритетному направлению электронных библиотек в «Программе работ по ИКТ на 2007-2008 годы» сформулированы следующим образом [9, стр. 36]:

«Создание новых подходов к сохранению информации и представлению знаний человека в цифровой среде на основе перспективных технологий по управлению динамически изменяемыми большими объемами данных, гарантирующих сохранение цифрового контента, выявление и **экспликацию эволюции его семантики**».

В приведенной формулировке говорится о «семантике цифрового контента», но, при этом, отсутствует явное или контекстное определение содержание этого словосочетания.

Проблема представления знаний упоминается еще в одном приоритетном направлении «Перспективные ИКТ» в теме «ИКТ долговременного применения». Ключевые положения этой темы, имеющие непосред-

025-Zatsman:_Layout 1 18.05.2008 15:11 Page 166

ственное отношение к проблеме представления знаний, сформулированы следующим образом [9, стр. 63]:

«Разработать новые подходы к представлению и сохранению знаний, ориентированные на долговременный и безотказный к ним доступ в условиях **локальной генерации концептов**, их интеграции, а также глобального использования систем представления и сохранения знаний с учетом контекста и временной эволюции систем. При этом должна быть обеспечена **долговременная устойчивость** систем представления и сохранения знаний в условиях многообразия их использования и **эволюции семантики во времени**».

Таким образом, кроме выявления и экспликации эволюции «семантики цифрового контента» ставится задача учета «локальности генерации концептов», обеспечения «долговременной устойчивости систем представления и сохранения знаний» и «интеграции концептов», но опять нигде не определено смысловое содержание используемых словосочетаний. Отметим, что формулировки «Программы работ по ИКТ на 2007-2008 годы» являются весьма лаконичными, что является следствием жанра этого документа. Иногда смысл целого ряда ключевых положений трудно понять из контекста этого программного документа. Поэтому потенциальные заявители проектов нередко вынуждены обращаться в Директорат 7-й Рамочной программы ЕС с просьбой пояснить смысл положений этой программы.

Что касается проблемы представления знаний, то более подробное современное ее описание можно найти в трудах семинара «Knowledge Anywhere Anytime: “The Social Life of Knowledge”», который состоялся 29-30 апреля 2004 года в Брюсселе [10]. Материалы этого семинара в редуцированном виде использовались при формировании «Программы работ по ИКТ».

Ключевые направления исследований по проблеме представления знаний

В материалах семинара отмечается, что исследование процессов генерации знаний и образования конвенциональных систем знаний, а также связанных с ними процессов, является актуальной проблемой, которая остается во многом нерешенной. Одновременно фиксируется тот факт, что содержание самой этой проблемы со временем эволюционирует. Участники семинара определили четыре актуальных направления исследований в рамках этой эволюционирующей во времени проблемы.

Задачей первого направления является формирование научного понимания того, как знания появляются, каким образом на этот процесс и его результаты влияет совместная деятельность, как формируются конвенциональные системы знаний. Одна из задач этого направления заключается в определении и фиксировании различий в понимании идентичных информационных объектов разными участниками совместной деятельности.

Задачей второго направления является исследование многообразия форм представления одних и тех же концептов, одной и той же системы знаний. Кроме форм представления конвенциональных и стабильных концептов, предметом исследования являются формы представления личностных и локальных концептов, а также процессы возникновения и эволюции нестабильных концептов. В рамках этого направления предполагается выполнение исследований процессов образования конвенциональных концептов на основе личностных и локальных концептов.

Задачей третьего направления является создание нового поколения интеллектуальных систем, которые должны обеспечить «семантическую интероперабельность» пользователей этих систем в процессе совместной работы. Если использовать термины семиуровневой модели интероперабельности Р. Будденберга, то следует различать два уровня «семантической интероперабельности»: когнитивная интероперабельность (6 уровень семиуровневой модели) и доктринальная интероперабельность (7 уровень). В соответствии с определением Р. Будденберга реализация когнитивной интероперабельности в процессе совместной работы предполагает обеспечение согласованного понимания пользователями идентичных информационных объектов, являющихся формами представления концептов. Реализация доктринальной интероперабельности предполагает не только согласованное понимание пользователями информационных объектов, но и обеспечение принятия ими согласованных решений на основе идентичной информации. Естественно, что в интеллектуальных системах нового поколения должна обеспечивается и «технологическая интероперабельность» на первых пяти уровнях семиуровневой модели, в том числе, на уровнях разделяемых процедур, процессов и данных [11, 12].

В рамках этого направления, кроме создания методов и средств поддержки «семантической интероперабельности» в интеллектуальных системах, планируется исследовать вопросы выявления и экспликации основных стадий эволюции систем знаний, представленных в виде классификационных систем, тезаурусов и онтологий, в процессе совместной работы пользователей. При этом, не предполагается, что пользователи заранее будут владеть согласованной между ними системой терминов и единым пониманием принципов построения систем знаний.

Степень новизны интеллектуальных систем предлагается оценивать на основе их сравнения с системами «управления знаниями» (knowledge management systems), основанных на редукционистском подходе к трактовке

025-Zatsman:_Layout 1 18.05.2008 15:11 Page 167

знаний человека, в котором, как правило, не различаются личностные, локальные и конвенциональные концепты, стабильные и нестабильные концепты.

Задачей четвертого направления является исследование принципиальных возможностей и средств влияния на формирование новых систем знаний в процессе совместной деятельности. Речь идет о формировании систем, в первую очередь, ориентированных на удовлетворение технологических, экономических, образовательных, экологических и другим социально значимых потребностей общества. В рамках этого направления предлагается исследовать, какими видами перспективных ИКТ и до какой степени можно оказывать влияние на процесс генерации новых знаний в процессе совместной деятельности, отвечающих социально значимым потребностям и способствующих получению целевых результатов совместной деятельности. Именно возможность оказывать влияние на процесс формирования новых целевых знаний является, по мнению участников семинара «Knowledge Anywhere Anytime: "The Social Life of Knowledge"», отличительной чертой общества, основанного на знаниях.

Приведенный перечень из четырех направлений исследований говорит о том, что участники семинара существенно расширяют границы традиционной проблемы представления знаний, сформулированной в рамках редукционистского подхода. Основная идея предлагаемого расширения заключается в рассмотрении генерации и распространения знаний как управляемых процессов. При этом, акцентируется социальная природа и социальная обусловленность процессов генерации и распространения знаний, что существенно усложняет управление этими процессами на основе ИКТ. Отмечается, что социальная природа этих процессов не противоречит существованию «компьютерного знания» (machine knowledge), если признать, что этот термин используется для обозначения роли, играемой компьютерными системами в процессах генерации, интеграции и распространения знаний.

Определяя актуальные направления исследований по проблеме представления знаний, эксперты не стали предлагать согласованную систему терминов для описания новых концептуальных положений. Пока новые идеи только обсуждаются и определяются направления исследований, подобная ситуация, как правило, неизбежна, так как система терминов для их описания только начинает формироваться. Однако, начиная исследования в новых направлениях, необходимо пытаться предлагать новые термины для описания новых идей и/или уточнять содержание ранее введенных терминов.

Задачи представления знаний и термины для их описания

Как отмечалось в начале доклада, для более четкого описания во многом нового взгляда на проблему представления и сохранения знаний в цифровой среде, представленного в документах 7-й Рамочной программы ЕС, необходимо определить новые термины и отношения между ними. Сначала приведем описание системы терминов из работы [1].

Начнем со слова «*знания*», которое будем трактовать как результаты познавательной и креативной деятельности человека, носителем которых может быть только человек и в которых могут быть выделены отдельные «кванты» знаний (в программных документах 7-й Рамочной программы используется словосочетание «knowledge parts»).

Информацию определим как формы эксплицитного и отчужденного от человека представления его знаний, предназначенные для передачи, непосредственного сенсорного восприятия и понимания их другими людьми.

В процесах комплексного кодирования основное внимание будет уделяться тем «квантам» знаний, называемых *концептами*, которые являются элементарными единицами или сочетаниями элементарных единиц плана содержания, выражаемого в рамках некоторого естественного языка (в общем случае, в рамках той или иной знаковой системы). Все остальные «кванты» знаний человека, которые не являются концептами (то есть, не являются элементарными единицами или сочетаниями элементарных единиц плана содержания), будем называть *ментальными единицами* знаний человека.

Приведенное определение термина «концепты» предполагает, что они являются результатом процесса членения знаний человека на «кванты», которые могут быть выражены в рамках некоторой знаковой системы. Процесс членения неразрывно связан с процессом выражения знаний человека в сенсорно воспринимаемой и отчужденной от человека форме, например, в виде текста на естественном языке или в виде геоизображения на языке карты.

В системе знаний человека могут быть выделены несколько планов содержания в зависимости от того языка или той знаковой системы, которыми он пользуется для представления своих знаний в отчужденной форме. Элементарные единицы плана содержания, имеющие значение в рамках некоторой знаковой системы, будем называть *элементарными концептами*.

Здесь необходимо отметить разницу между элементарным концептом и значением слова, когда речь идет о естественном языке. В процессах комплексного кодирования научно-технических документов рассматривается

025-Zatsman:_Layout 1 18.05.2008 15:11 Page 168

только сигнификативный аспект значения. Экспрессивно-эмоциональные оценки и коннотации не рассматриваются, так как речь идет о представлении научно-технических знаний в виде научно-технических документов, а не в виде художественных произведений. Следовательно, термины «концепт» и «элементарный концепт» используются для обозначения только сигнификативной составляющей значения слова [13].

Термин «*знаковая информация*» определим как результаты процесса представления концептов человеком-генератором этих результатов в плане выражения сферы социальных коммуникаций в любой отчужденной форме, которая является сенсорно воспринимаемой другими участниками коммуникаций. Отметим, что при таком определении термин «знаковая информация» имеет отношение только к формам представления концептов, а введенный ранее термин «информация» - к формам представления любых «квантов» знаний, включая концепты.

Существительное «информация» является неисчисляемым, что не всегда удобно для описания процессов комплексного кодирования. Поэтому, определим следующие термины с использованием исчисляемых существительных. Представление в плане выражения элементарных концептов в виде сенсорно воспринимаемых форм будем называть *элементарными информационными объектами*. Сенсорно воспринимаемые формы представления любых концептов, включая элементарные, будем называть *информационными объектами*.

Отметим, что выделение в системе знаний человека нескольких планов содержания позволяет учесть различия в членении знаний человека в разных естественных языках и других знаковых системах [14].

Термин «*коды*» определим как компьютерные эквиваленты двоичных цифр (или их последовательностей), которые могут представлять собой намагниченность или ее отсутствие, наличие электрического тока или его отсутствие, способность к отражению света или ее отсутствие в цифровой среде [15, с. 86]. Двоичные цифры «0» и «1», о которых говорится в определении термина «коды» принадлежат сфере социальных коммуникаций, а их компьютерные эквиваленты - цифровой среде.

При описании процессов комплексного кодирования будем выделять среди всех возможных кодов цифровой среды три следующих категории:

- коды, соотнесенные с концептами знаний человека (*первая категория*),
- коды, соотнесенные с эксплицитными и отчужденными от человека формами представления концептов в плане выражения сферы социальных коммуникаций (*вторая категория*),
- коды, соотнесенные с материальными объектами и их свойствами, отношениями, ситуациями, состояниями, процессами, действиями, алгоритмами, программами и другими категориями денотатов (*третья категория*).

Например, для кодирования значений слов будут использоваться коды первой категории, для кодирования сочетаний литер алфавитных систем письма, составляющих слова, будут использоваться коды второй категории, а для кодирования предметов, обозначаемых этими словами, будут использоваться коды третьей категории.

Перечислим рассмотренные термины, разделив их на три части в зависимости от их природы (ментальная, социальная и цифровая) и выделив курсивом термины, определение которых не зависит от термина «знаковая система»:

- *знания*, концепты и элементарные концепты (ментальная сфера),
- *информация*, знаковая информация, информационные объекты, элементарные информационные объекты (сфера социальных коммуникаций),
- *коды*, категории кодов (цифровая среда).

Эти термины являются основой для построения дефиниций стабильных, личностных, локальных и нестабильных концептов. Отметим, что смысл термина «концепт» существенно зависит от дефиниции термина «знаковая система», так как любой концепт определяется в рамках той или иной знаковой системы. Однако по определению знака и знаковой системы две стороны знака (форма как материально выраженная его составляющая и значение как его идеальная составляющая), будучи поставлены в отношение *постоянной связи*, опосредованной сознанием, составляют *устойчивое* единство, которое посредством сенсорно воспринимаемой формы знака репрезентирует *конвенционально* приданное ему значение [16].

Таким образом, приведенное энциклопедическое определение термина «концепты» предполагает, что они являются и конвенциональными, и стабильными. Чтобы определить личностные и локальные концепты, сначала введем понятия авторского и локального знаков. Авторский знак отличается от традиционного семиотического знака тем, что две стороны авторского знака – форма и значение знака – могут находиться в отношении *временной связи*, опосредованной сознанием **одного человека**, составлять *нестабильное* единство, которое посредством сенсорно воспринимаемой формы знака этим человеком репрезентирует *персонально* приданное ему значение в течение некоторого периода времени. Локальный знак отличается от авторского тем, что две стороны локального знака могут находиться в отношении *временной связи*, опосредованной сознанием каждого из участников совместной деятельности, то есть используют и согласованно понимают локальные знаки, как минимум, **два человека.**

025-Zatsman:_Layout 1 18.05.2008 15:11 Page 169

Авторские и локальные знаки не возникают и тем более не функционируют раздельно. В своей совокупности с традиционными знаками они образовывают единую знаковую систему. Иначе говоря, авторские и локальные знаки означиваются в совокупности с традиционными знаками, являющимися составными элементами естественного языка или невербальной знаковой системы[2].

Используя определения авторского и локального знаков, определим элементарный личностный концепт как значение авторского знака, а личностный концепт как значение выражения на естественном языке, содержащего хотя бы один авторский знак, либо новое значение выражения на естественном языке без авторских знаков, новый смысл которого определен автором в явном виде.

Определим элементарный локальный концепт как значение локального знака, а локальный концепт как значение выражения на естественном языке, содержащего хотя бы один локальный знак, либо новое значение выражения на естественном языке без локальных знаков, новый смысл которого определен в явном виде и согласованно понимается участниками совместной деятельности.

В заключение раздела сформулируем подход к определению границ между стабильными (lasting concepts) и нестабильными концептами (volatile concepts), коды значений и/или форм которых хранятся в ИЭБ. За основу возьмем принцип разделения МПК8 на рубрики базового и расширенного уровней. Как отмечалось в начале доклада, рубрики базового уровня, выражающие стабильные концепты, могут пересматриваться не чаще, чем раз в три года, а рубрики расширенного уровня могут пересматриваться значительно чаще.

Будем говорить, что в рамках некоторой институциональной системы существует граница между стабильными и нестабильными концептами, если в институциональной системе классификации и/или тезаурусе, которые используются для представления знаний в ИЭБ, разделены стабильные и нестабильные рубрики системы классификации (стабильные и нестабильные дескрипторы тезауруса).

Заключение

Новые грани проблемы представления знаний в цифровой среде, рассмотренные в докладе, сформировались во многом под влиянием ряда институциональных факторов. Наиболее четко их влияние проявляется сейчас в патентной сфере, в рамках которой внесены существенные изменения в структуру и регламент ведения институциональной классификационной системы - МПК. Два ключевых институциональных фактора 1) новая структура МПК, в который впервые выделены базовый и расширенный уровни, и 2) ежеквартальный график заседаний Специального подкомитета ВОИС по пересмотру расширенного уровня МПК, с одной стороны, нормативно разделили стабильные и нестабильные концепты, с другой стороны, зафиксировали основные стадии формирования конвенциональных концептов на основе авторских и локальных концептов.

Приведенный перечень актуальных направлений исследований, в которых нашла отражение эволюционирующая проблема представления и сохранения знаний в цифровой среде, позволяет утверждать о необходимости развития соответствующей терминосистемы. Один из возможных вариантов развития, рассмотренный в докладе, основан на понятиях авторского и локального знаков. Кроме понятий авторского и локальных знаков, использовалась система базовых терминов, предложенная ранее. Основным критерием, определяющим предлагаемое направление развития терминосистемы, является стремление к моносемичности новых терминов, расширяющих систему базовых терминов из работы [1].

Список литературы

1. Зацман И.М. Концептуальный поиск и качество информации.– М.: Наука, 2003.

2. Введение в МПК8 - http://www.fips.ru/ipc8/intro/mpk8.htm.

3. О новом порядке пересмотра и реализации МПК расширенного уровня - http://www.fips.ru/russite/classificators/new.htm.

4. Зацман И.М., Кожунова О.С. Семантический словарь системы информационного мониторинга в сфере науки: задачи и функции // Системы и средства информатики. Вып. 17. - М.: Наука, 2007. - С. 124-141.

5. Кожунова О.С., Зацман И.М. Прагматические аспекты создания семантического словаря терминов информационного мониторинга // Труды международной конференции Диалог-2007 «Компьютерная лингвистика и интеллектуальные технологии». - М.: Изд-во РГГУ, 2007. - С. 278-285.

6. Зацман И.М. Вербально-образное представление знаний в электронных библиотеках (Часть II) // Научно-техническая информация (серия 2 «Информационные процессы и системы»). - 2001. No 12. - С. 10-17.

[2] Приведенные определения авторского и локального знаков основаны на развитии положений из работы [16].

025-Zatsman:_Layout 1 18.05.2008 15:11 Page 170

7. Decision No 1982/2006/EC of the European Parliament and of the Council of 18 December 2006 concerning the Seventh Framework Programme of the European Community for research, technological development and demonstration activities (2007-2013) // Official Journal of the European Union L412 30.12.2006. – P. 1-41.

8. CORDIS ICT Programme Home - http://cordis.europa.eu/fp7/ict/programme/ home_en.html (состояние страницы на 27.07.2007).

9. ICT FP7 Work Programme - ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/ict-wp-2007-08_en.pdf (состояние файла на 27.07.2007).

10. FP7 Exploratory Workshop 4 «Knowledge Anywhere Anytime» - http://cordis.europa.eu/ist/directorate_f/f_ws4.htm (состояние страницы на 04.02.2008).

11. Buddenberg R. Toward an Interoperability Reference Model - http://web1.nps.navy.mil/~budden/lecture.notes/interop_RM.html (состояние страницы на 12.02.2008).

12. Buddenberg R. FORCENet: We've been here before - http://web1.nps.navy.mil/~budden/lecture.notes/it_arch/large_info_systems.html (состояние страницы на 12.02.2008).

13. Гак В.Г. Лексическое значение слова // Большой энциклопедический словарь «Языкознание». – М.: Большая российская энциклопедия, 1998. – С. 261-263.

14. Vossen P. (ed.) EuroWordNet General Document (Version 3) (URL: http://www.illc.uva.nl/EuroWordNet/docs/GeneralDoc).

15. McArthur D. Information, its forms and functions: The elements of semiology. –Lewinton: The Edwin Mellen Press, Ltd., 1997.

16. Уфимцева А.А. Знак языковой // Большой энциклопедический словарь «Языкознание». – М.: Большая российская энциклопедия, 1998. – С. 167.

026-Iomdin_B:_Layout 1 18.05.2008 14:51 Page 171

ИДЕЯ ОДНОИМЕННОСТИ В РУССКОМ ЯЗЫКЕ[1]
THE IDEA OF MATCHING NAMES IN RUSSIAN

Иомдин Б.Л. *(iomdin@ruslang.ru)*
Институт русского языка им. В. В. Виноградова РАН

Работа посвящена метаязыковой лексике, описывающей определенные соотношения названий: языковым единицам одноименный и так и наз(ы)вать(ся) и некоторым другим. Интерпретация единиц такого рода – трудная задача для автоматической обработки текстов. Предложены толкования и выделено несколько ключевых смыслов.

Введение

Настоящая работа посвящена русской метаязыковой лексике с нетривиальной сферой действия, описывающей соотношения названий друг с другом и с именуемыми объектами. Конкретнее, нас будут интересовать в первую очередь языковые единицы *одноименный* (со своими производными) и *так и наз(ы)вать(ся)*, а также, в меньшей степени, и некоторые другие.

1. Одноименный

Прежде всего, обратимся к словарям. Вот как представлена лексема *одноименный* в нескольких традиционных толковых словарях русского языка:

• ‘Носящий то же имя, название, что и другой’: *Одноименные города* (СУш);

• ‘Носящий то же самое имя, название’: *Обломов — герой одноименного романа И. Гончарова* (МАС);

• ‘Носящий то же имя, название’: *Одноименные поселки. Фильм по одноименному роману* (СОШ).

Легко видеть, что в примерах МАС и СУш лексема *одноименный* употребляется по-разному. В самом деле, словосочетания типа *одноименные рестораны* неоднозначны; сравним фразы *В одноименных ресторанах часто бывает разное меню* и *Иван был в Токио и Пекине, а Петр – лишь в одноименных ресторанах*: в первой фразе *одноименный* употреблено так, как в СУш, а во второй – так, как в МАС[2]. Примеры в СОШ представляют оба типа употребления, которым, однако, дано общее толкование.

Итак, цитированные словари явно неполно описывают слово *одноименный*. Нам представляется, что оно заслуживает более подробного изучения и более точного описания. Ниже мы предлагаем выделять у него три типа употребления, по-видимому, соответствующие трем разным лексическим значениям.

1.1. Анафора. Основная лексема слова *одноименный* – анафорическая: ‘такой, название которого совпадает с ранее упомянутым названием другого объекта’[3]. Эта лексема в какой-то мере близка другим анафорическим прилагательным типа *названный*, *(выше)указанный*, *(выше)упомянутый*, однако она указывает на совпадение названий, т.е. имен объектов, а не самих объектов. Ср. (1)–(2):

(1) *Недавно побывал на премьере фильма «Звезда» по одноименной повести Эммануила Казакевича* («Известия», 2002.06.13)

(2) *Можно создать «Радио хороших новостей», выпускать одноименную ежедневную газету, аналитический еженедельник того же направления* («Совершенно секретно», 2003.04.08)

Немного более сложная ситуация представлена в случаях, когда эксплицитное название отсутствует, а антецедентом лексемы *одноименный* является какой-то фрагмент предтекста. Ср. (3)–(4):

[1] Работа выполнена при финансовой поддержке Программы фундаментальных исследований отделения историко-филологических наук РАН «Русская культура в мировой истории», фонда РГНФ (грант №06-04-00289а) и гранта Президента РФ для поддержки научных исследований, проводимых ведущими научными школами РФ №НШ-3205.2008.6. В работе использованы примеры из корпуса текстов Сектора теоретической семантики ИРЯ им. В. В. Виноградова РАН и Национального корпуса русского языка (www.ruscorpora.ru).

[2] Два аналогичных значения выделяются у слова *подобный* в [Богуславская 2004: 810]. Ср. *подобные треугольники* (‘треугольники подобны друг другу’) vs. *В подобное место я и шагу не ступлю* (‘место подобно другому месту, упомянутому ранее’).

[3] Высказывания, где *одноименный* сочетается не с именами объектов, а со словами типа *имя* <*название*>, представляются нам неправильными, хотя они также иногда встречаются в текстах; ср. *Эта картина – продолжение ленты с одноименным названием, посвященной теме абортов* («Спецназ России», 2003.02.15); *Нападение на Богатырева переполошило весь «Аэропорт», то есть писателей, которые жили у станции метро с одноименным названием* (В. Войнович, Дело № 34840).

026-Iomdin_B:_Layout 1 18.05.2008 14:51 Page 172

(3) *Принято считать, что лучшие программы по культуре сосредоточены на одноименном российском канале* («Вестник США», 2003.09.03)

(4) *Что такое хорошо и что такое плохо, Женя усвоила примерно в те же годы, когда впервые прочитала одноименное стихотворение Маяковского* (А. Берсенева, Возраст третьей любви)

Этот фрагмент может отстоять от лексемы *одноименный* достаточно далеко, что создает дополнительные трудности при попытке автоматического определения ее антецедента. Ср. (5):

(5) *Что такое в точности мартовские иды, никто из нас не помнил. Мне почему-то казалось, что это праздники, и я выпросила название для заметки, в которой никак нельзя обойтись без грядущего 8 марта. Пришлось взяться за дело: я просмотрела несколько энциклопедий, перерыла Интернет, позвонила знакомым, которых сто лет не видела, перелистала одноименный роман Торнтона Уайлдера и много смеялась* («Домовой», 2002.03.04)

Ситуации, когда антецедент лексемы *одноименный* не является эксплицитно выделенным названием, вообще очень трудны для автоматической интерпретации, особенно при наличии в предтексте других кандидатов на это место, оформленных как названия. Ср. (6):

(6) *Кабаниха из «Грозы» Островского, Простакова из «Недоросля», мать из одноименной повести Горького, пушкинская Арина Родионовна и другие роли были сыграны ею* («Восточно-Сибирская правда», 2003.06.14).

Вот пример еще более сложной отсылки (без знаний о мире весьма трудно определить, что в (7) речь идет о саде «Эрмитаж»):

(7) *Нищая богема сидела в «Гагарин-пати», «Мобиле» или «Эрмитаже». Первый клуб – это павильон «Космос» на ВВЦ, второй – велотрек в Крылатском, третий – театр в одноименном саду* («Мир & Дом. City», 2003)

Некоторые высказывания и вовсе невозможно интерпретировать без привлечения экстралингвистической информации. Ср. (8):

(8) *А директора санатория «Дубовая роща» Игоря Туманова все время упорно причисляют к братьям Тумановым, возглавлявшим известную в середине 90-х одноименную преступную группировку* («Дело» (Самара), 2002.03.19)

Как называлась группировка: «братья Тумановы», «тумановские», «тумановская группировка», «Дубовая роща»?.. Можно что-то предполагать лишь с использованием нетривиального соображения о том, что преступная группировка вряд ли носила официальное название (тем более такое романтическое, как «Дубовая роща»).

Кроме того, возможна также катафора, когда лексема *одноименный* относится к фрагменту последующего текста. Ср. (9):

(9) *Снятый по мотивам одноименного комикса «Каратель» несмотря ни на что впечатляет* («Хулиган», 2004)

Наконец, не исключена и комбинация двух последних типов употребления *одноименный* – катафора без эксплицитного упоминания названия. Ср. (10)–(11):

(10) *Центр одноименного района, Стерлитамак имеет около 26 тысяч жителей* (Н. Покровский, По Белой)

(11) *Основным ценителем спортивных качеств конкурсанток, по всей видимости, должен был выступать генеральный директор шейпинг-центра «Вера» Вячеслав Белешин. Оценивать прически и макияж – директор одноименного салона Сергей Бухтияров* («Дело» (Самара), 2002.06.17)

Правильно интерпретировать последний пример (в действительности салон называется «Сергей Бухтияров», по имени директора) достаточно трудно.

1.2. Совпадение названий. Значение ‘имеющий одинаковые названия’, представленное примерами из СУш и СОШ (*одноименные города*, *одноименные поселки*), достаточно редко встречается в текстах. Ср. (12):

(12) *Скоро выяснилось, что в округе две одноименных деревни, эта и разыскиваемая* (Б. Пастернак, Доктор Живаго).

Отметим, что в этом значении *одноименный*, по-видимому, сочетается лишь с именными группами в форме МН. Ср. похожее поведение форм МН слов *друг*, *товарищ* и *приятель*, описанное в [Урысон 2004]: «Формы МН **друзья**, **товарищи** и **приятели** двузначны: они могут обозначать и группу людей, каждый из которых в отдельности является **другом** (**товарищем**, **приятелем**) субъекта, и группу людей, состоящую из пары, тройки и т. п. **друзей** (**товарищей**, **приятелей**). Ср. *Приходили его **друзья** <**товарищи**, **приятели**>* и *Друзья <**товарищи**, **приятели**>* решили ехать вместе. Ср., в противоположность этому, форму МН *знакомые*, у которой есть только первое из указанных значений» [Урысон 2004: 298].

Интересно, что, вопреки своей внутренней форме, прилагательное *одноименный* не может называть людей, имеющих одинаковые имена (или фамилии): ср. неправильность **В классе три одноименных мальчика:*

026-Iomdin_B:_Layout 1 18.05.2008 14:51 Page 173

Андрей Иншаков, Андрей Мякутин и Андрей Шулимов; **Маяковский, Набоков и Познер – одноименные люди*. Для таких случаев используются специальные лексемы *тезка* и *однофамилец*.

1.3. Соответствие. Близкое к предыдущему значение ‘соответствующий’, не представленное в известных нам толковых словарях, используется в основном в специальных текстах. Ср. *Одноименные полюса отталкиваются*, а также (13):

(13) *Склеив между собой некоторые многоугольники по одноименным сторонам, получаем другую, хорошо известную, крестообразную развертку куба* (Н. Долбилин, Жемчужины теории многогранников).

В отличие от предыдущей лексемы слова *одноименный*, эта лексема может сочетаться и с именными группами в форме ЕД – в конструкциях типа *X, одноименный с Y*; ср. (14):

(14) *Рука, одноименная с ногой, выставленной вперед, может быть для увеличения трения обвита одним витком веревки* (Л. Гутман, С. Ходакевич, И. Антонович, Техника альпинизма)

Эта лексема имеет дериват *одноименно* (ср. *намагничивать одноименно*) и антоним *разноименный*; ср. *Разноименные полюса притягиваются*, а также (15):

(15) *Рубяще-режущие или колющие движения по диагонали сверху вниз от одноименного и разноименного плеча и снизу вверх от одноименного и разноименного бока* («Боевое искусство планеты», 2004)

***

Вторая часть доклада посвящена устойчивому словосочетанию *так и наз(ы)вать(ся)*. От лексемы *одноименный* эта близкая на первый взгляд языковая единица существенно отличается.

Во-первых, она имеет другую структуру: если первая употребляется в сочетаниях типа *одноименный X*, где не упоминается название *X*-а (его надо искать в предшествующем тексте, см. выше), то вторая – в сочетаниях типа *X так и называется: «Y»*, где название явно указывается в последующем тексте. Ср. (16) и (17) – характерную пару не вполне нормативных, как представляется, примеров, где это правило нарушается:

(16) *Владислав рисует его в любую погоду, любуясь, например, достаточно незатейливым городским пейзажем, либо изумительными Пятницкими воротами в солнечный день. Восторг этот запечатлен в одноименной работе «Пятницкие ворота», исполненной им маслом в достаточно пастозной экспрессивной технике* («Жизнь национальностей», 2001)

(17) *Юрий Аввакумов увидел московские дома снизу вверх и так и назвал свою серию* («Известия», 2002.01.24)

В (16) название работы «Пятницкие ворота» лишнее (оно и так определяется словом *одноименный*); в (17), напротив, не становится ясным, как называется серия фотографий («Дома снизу вверх»? «Московские дома снизу вверх»? «Московские дома снизу вверх глазами Юрия Аввакумова»?...)

Во-вторых, в случае *одноименный* речь идет о буквальном совпадении названий (может быть, с точностью до частеречной принадлежности: ср. (10), где город называется Стерлитамак, а район – Стерлитамакский). Что же касается второй обсуждаемой языковой единицы, то описываемая ею связь между названием и именуемым объектом может быть как прямой, так и весьма условной, в зависимости от типа употреблений.

Основные типы употреблений *так и наз(ы)вать(ся)* обсуждаются в следующем разделе.

2. Так и наз(ы)вать(ся)

Нам не удалось найти в литературе лексикографического описания этой языковой единицы. Ее употребления достаточно разнообразны, но можно выделить несколько типичных случаев.

Прежде всего, интересно, что типы употреблений *так и наз(ы)вать(ся)* можно разделить на два больших блока с почти противоположными смыслами. В первый блок мы помещаем употребления, маркирующие соответствие названия ожиданиям говорящего или слущающего, а во второй блок – употребления, маркирующие противоречие названия таким ожиданиям.

2.1. Естественные названия

2.1.1. Простое название. ‘Название самым естественным образом связано с именуемым объектом, описывает его главное свойство’. Ср. (18)–(19):

(18) *Деду Морозу приходит много писем. Пришлось даже построить свою почту, она так и называется «Почта Деда Мороза»* («Мурзилка», № 12, 2002)

(19) *Толга – местечко на левом берегу Волги, выше Ярославля. Там был монастырь, он так и назывался – Толжский монастырь* (В. Розов, Удивление перед жизнью)

2.1.2. Мотивированное название. ‘Название объекта соответствует некоторому его свойству, ранее указанному в тексте’. Ср. (20)–(22):

026-Iomdin_B:_Layout 1 18.05.2008 14:51 Page 174

(20) *Интересно, как она называется? На ежика похожа, такая же круглая и в колючках. – Ты угадала, эта рыба так и называется – «рыба-еж», – пояснил Семен Семенович* (В. Постников, Карандаш и Самоделкин в стране фараонов)

(21) *Кроме того, вам должно быть известно, сударыня, что жемчуг рождается на внутренних створках особых раковин, которые так и называются «жемчужницы»* (В. Катаев, Жемчужина)

(22) *Как я был кострожогом, я описал в стихотворении. Оно так и называется – «Кострожоги»* (А. Жигулин, Черные камни)

При этом обоснованность названия может быть весьма условной, в предельном случае – понятной лишь автору текста. Ср. (23):

(23) *Для задуманной серии они позвали талантливых людей – Давыдова, Трифонова, Окуджаву. Книга Булата Окуджавы о Павле Пестеле так и называлась – «Глоток свободы»* («Вестник США», 2003.10.01).

Стиль такого рода, злоупотребляющий указанным выражением, пародирует Сергей Довлатов:

(24) *Я написал передачу о камнерезах. Передача так и называлась – «Живые камни»* (С. Довлатов, Соло на ундервуде)

Интересно, что у одного и того же объекта могут быть признаны главными совершенно разные свойства, мотивирующие совершенно разные названия. Ср. характерную пару примеров (25)–(26):

(25) *В 1901 году жители Лондона могли наблюдать интересную картину: к нужному дому две лошади подвозили громадную установку, рабочие разматывали шланги длиной в 250 метров, протягивали в комнаты трубы и начинали отсасывать пыль. Агрегат так и назывался: пылесос* (В. Быков, О. Деркач, Книга века)

(26) *Американцы были так поражены самой возможностью всасывания пыли неким механизмом, что ни о каком дизайне речи не шло. Например, модель 1908 года W.H.Hoover Company, сделанная из консервной жести и дерева, так и называлась – «Жестяная модель»* («Эксперт: Вещь», 2001)

2.1.3. Название, подтверждающее мысль говорящего. ‘Название объекта соответствует тому, что о нем или о чем-то с ним связанным утверждает говорящий’. Ср. (27)–(28):

(27) *Натэлла обладала потрясающим качеством – она сумела полностью обеспечить бытовой комфорт для этих двух мужчин, создать очаг и уют. У них на кухне было особое пространство, которое так и называлось – очаг* (С. Спивакова, Не всё)

(28) *Главная статья доходов города – налоги от газовиков; единственная современная улица в Мышкине так и называется – улица Газовиков* («Сельская новь», 2003)

2.2. Неожиданные названия

2.2.1. Родовое название. ‘В качестве названия объекта используется его родовое имя, хотя обычно у таких объектов существуют специальные названия’. Ср. (29)–(30):

(29) *К юбилею писателя вышел двухтомник его прозаических произведений, который так и называется «Проза»* («64 – Шахматное обозрение», 2004)

(30) *Мы, например, обедали в Столовой. Так и называется – столовая* (С. Яковлев, Очередной отчет об очередной поездке в очередные горы)

Необходимость в специальных названиях у некоторых объектов, отличных от простого описания их сути, часто ощущается говорящими. Ср. оценочное словосочетание *немудрящее название*, а также (31):

(31) *Все науки как-нибудь называются: химия, кибернетика, этимология, сурдопедагогика. А как называется наука воспитания? Ну? – Так и называется: теория воспитания. – Теория! – Каштанов хмыкнул. – Теория – это теория, а наука – это наука* (С. Соловейчик, Ватага «Семь ветров»)

2.2.2. Чересчур прямое название. ‘Название объекта не скрывает отношения к нему субъекта именования’. Ср. (32):

(32) *Конечно, есть и такие ребята, которые воруют, но зато их так и называют: шпана* (В. Белов, Мальчики)

Обычно такое название выражает отрицательное отношение субъекта (именующего) к объекту (именуемому):

(33) *Восточной Пруссии в Баварии не любили, так и называли «свинские пруссаки»* (И. Сабурова, О нас)

(34) *Это была «элита». А четверть была из микрорайона – чтоб не вязались… разные там роно, гороно… Их так и называли – «микрорайонские». Сами учителя и называли. И относились к ним соответственно* (В. Белоусова, Жил на свете рыцарь бедный)

2.2.3. Неуместное название. ‘Название действительно именно таково, как произнес ранее говорящий, несмотря на его кажущуюся непригодность или неуместность’. Ср. (35)–(37):

(35) *Тебе некто Колька звонил. Так и назвался: Колька* (Л. Карелин, Головокружение)

(36) *Физический смысл я здесь не совсем понимаю, но все равно – здорово получаются эти пузыри. А что? Так и назову: пузыри. Нет, наверное, лучше «полости». Полости Малянова. «М-полости»*

026-Iomdin_B:_Layout 1 18.05.2008 14:51 Page 175

(А. Стругацкий, Б. Стругацкий, За миллиард лет до конца света)

(37) *Крылов забирался туда и, приоткрыв низенькую дверцу, слушал, как теоретики «трепались». Свои семинары они так и называли «треп». Официальное наименование «семинар» совсем не подходило к этому шумному сборищу, где все серьезное перемежалось шутками и, пока писали формулы, рассказывали анекдоты* (Д. Гранин, Иду на грозу)

3. Ключевые смыслы и другие языковые единицы с близкими значениями

В предложенных выше толкованиях можно выделить несколько общих смыслов, которые содержатся и в значениях многих других русских языковых единиц. Прежде всего это идеи совпадения, соответствия и простоты.

3.1. Совпадение. Совпадение названий, описываемых лексемой *одноименный*, не является случайным: второе из них появилось благодаря первому; ср. также языковые единицы *в честь, имени (кого-л.)*.

Высказывания, где речь идет об «омонимии» названий, представляются комичными или игровыми; ср. (38), где названия рыбы и города явно не произведены друг от друга:

(38) *В Энциклопедическом словаре о Петушках две строчки: Петушки, г. (с 1965) во Владимирской обл., на р. Клязьма. Ж.-д. ст. Текст. пр-сть. Рядом об одноименной рыбе из семейства лабиринтовых. В три раза больше. Словом, глушь* («Совершенно секретно», 2003.03.02)

Высказывания, где названием-источником является название второго, а не первого из упомянутых объектов, также используются лишь для создания юмористического эффекта, ср. (39):

(39) *Возбудитель болезни – палочки Коха, микробы, открытые одноименным ученым 102 года назад* («Столица», 1997.04.15)

Для *так и наз(ы)вать(ся)* идея совпадения также важна, однако здесь могут описываться и случайные совпадения. Ср.

(40) *Поразительным было то, что встретились мы с этим негром на мосту в Париже, а* [написанное раньше] *стихотворение так и называлось «На мосту» и тоже было о Париже* (Е. Евтушенко, «Волчий паспорт»)

Идея совпадения выступает на первый план в значении таких языковых единиц, как *тезка*, *однофамилец*, а также *говорящая фамилия* <*говорящее название*>; ср. (41)–(42):

(41) *Бригадному генералу с говорящей армейской фамилией Плиев и просто генералам с продовольственными микояновскими фамилиями Гречко и Гарбуз был устроен полноценный нагоняй* (А. Архангельский, Послание к Тимофею)

(42) *Таковая [мастерская] нашлась на окраине Москвы, на улице с говорящим названием Монтажная* («Автопилот», 2002.10.15)[4]

3.2. Соответствие. Впрочем, для русского языка в первую очередь важна идея неслучайности, мотивированности названий (а возможно, и влияния их на свойства именуемых объектов: ср., например, известную шуточную фразу *Как вы лодку назовете, так она и поплывет*). Ср. характерные примеры со словами *недаром*, *неспроста*:

(43) *В этом Бернес сильно напоминал мне моего папу. Недаром их обоих звали прекрасным мужским именем – Марк* (Л. Гурченко, Аплодисменты)

(44) – *Слово душа неспроста с дыханием созвучно, – дедушка пригладил растрепавшиеся седые вихры. – Ветер, он тоже дыхание* (М. Елизаров, Pasternak)

Эта идея часто эксплуатируется для создания юмористического эффекта. Ср. (45)–(46):

(45) *При выходе один на один вратарь врезал мне по ноге. Было так больно, что, казалось, ногу отпилили. Неспроста, видно, по-польски футбол называется «пилка ножна»* («Столица», 1997.08.26)

(46) *Отвлеченные науки, которыми набита ваша молодая голова, потому и называются отвлеченными, что они отвлекают ваш ум от очевидности* (А. Чехов, Дуэль)[5]

С этим же связана распространенность явления «народной этимологии», когда люди стремятся найти

[4] Хотя говорящие названия чаще всего являются результатом совпадения, прямое противоречие названия объекта его свойствам также иногда отмечается говорящими; ср. У*лица Мясников состояла, как его и предупреждали, главным образом из разнокалиберных — наоборот — рыбных ресторанов* (Weekend, № 8(54), 07.03.2008).

[5] Ср. также каламбур А. Милна: «It‘s—I wondered—It‘s only—Rabbit, I suppose you don‘t know, What does the North Pole look like?» <…> «I suppose it’s just a pole stuck in the ground?» «Sure to be a pole,» said Rabbit, «because of calling it a pole, and if it’s a pole, well, I should think it would be sticking in the ground, shouldn’t you, because there’d be nowhere else to stick it» (A.A.Milne, Winnie-the-Pooh), идея которого отчасти сохранена и в пересказе Б. Заходера: «Конечно, там есть ось, и, конечно, она воткнута в землю, потому что больше же ее некуда воткнуть, да к тому же она так и называется: «земляная»».

026-Iomdin_B:_Layout 1 18.05.2008 14:51 Page 176

мотивацию непонятного названия, в особенности заимствования. Ср. эксплуатацию этого явления в (47):

(47) *Сюда гнали каторжан в кандалах. До одной станции. Там кандалы снимали, потому что бежать уже было дальше некуда. Станция поэтому называется «Кандалакша»* (Л. Вертинская, Синяя птица любви)

Идея соответствия названий и именуемых объектов настолько сильна, что порой эти сущности смешиваются. Ср. характерное употребление лексемы *сам* в (48):

(48) *В моей семье Пасха – праздник святой. <…> В этот светлый день мы красим яйца (причем не меньше ста штук), печем куличики (и тоже не меньше сотни) и готовим саму пасху*[6] (Интернет-издание Дни.Ру).

Ср. также распространенность высказываний вроде *X потому и называется X*, в которых название X на самом деле никак не мотивировано указанным свойством:

(49) *Трутни потому и называются трутнями, что они в жизни ничего не делают или делают только то, что им захочется! А ведь это и есть наша с тобой мечта, Малинин!* (В. Медведев, Баранкин, будь человеком!)

(50) *Не заставший ту эпоху скажет недоуменно: но работа потому и называется общественной, что за нее не платят, вот бы и отказывались* (Е. Рубин, Пан или пропал. Жизнеописание)

Из самого названия *общественная работа* не вытекает, что такая работа не оплачивается, а слово *трутень* вообще не имеет (прозрачной) внутренней формы; тем не менее связь означаемого и означающего настолько прочна, что и эти наименования кажутся мотивированными[7].

3.3. Простота. Смысл ‘простота’ в русском языке, как известно, очень важен и очень сложен. Нам показалось интересным сравнить выделенные нами значения языковой единицы *так и наз(ы)вать(ся)* с семантической структурой прилагательного *простой*, описанию которой посвящена большая работа [Бабаева 2006]. В этой работе выделяется шесть основных идей, лежащих в основе разных лексем слова *простой*; здесь мы лишь коротко перечислим их, иллюстрируя примерами Е. Э. Бабаевой:

Идея стандартности: *Нарисуй на бумаге простой кружок*; *простая история*

Идея иерархии: *простой карандаш, простой народ*

Идея базовости: *простой механизм <организм>*

Идея ограничительности: *простое любопытство*; *простая халатность*

Идея естественности: *простая кожа, простой помол*, *простодушие*

Идея прямизны: *простой путь* [Бабаева 2006].

Добавим, что идея прямизны, отсутствия хитростей, недомолвок и эвфемизации содержится не только в описанных выше употреблениях *так и наз(ы)вать(ся)*, но и в близком к нему выражении *прямо называть* или комбинации обоих выражений. Ср. (51)–(52):

(51) *И вообще, главная героиня прямо названа в романе ведьмой* («Лебедь» (Бостон), 2003.10.26)

(52) *В 1929 году в СССР получили первый металлокерамический твердый сплав. Его так прямо и назвали – победит, обозначая действие, которое он произведет над любой твердой поверхностью* (В. Быков, О. Деркач, Книга века).

***

Выделение оппозиций (1) «случайное совпадение vs. неслучайное совпадение» и (2) «совпадение имен объектов vs. соответствие имен объектов и свойств этих объектов» позволяет разделить рассмотренную лексику на четыре класса. В первый класс слов (описывающий случайное совпадение имен объектов) входят лексемы *одноименный* (во втором значении), *тезка, однофамилец, омоним*[8]. Во второй (неслучайно совпадающие имена объектов) – языковые единицы типа *одноименный* (в первом значении), *в честь* и *имени (кого-л.)*. В третий (имена объектов, случайно совпадающие с их свойствами) – употребления типа *говорящая фамилия* (если речь идет о реальных людях, а не о литературных героях, которым говорящие фамилии обычно придумываются специально). Наконец, в четвертый класс (имена объектов, соответствующие их свойствам) входит выражение *так и наз(ы)вать(ся)* (во всяком случае, первый блок его употреблений). Насколько такая классификация универсальна, можно ли найти языковые единицы всех четырех классов и в других языках – вопрос отдельного исследования.

[6] Поясним, что здесь речь идет о творожном блюде, которое обычно готовят к празднику Пасхи и которое так и называется.

[7] Высказывания такого типа сближаются с конструкциями вида *X – это X*, описанными в [Шмелев 1996: 185]: *Ложка – это ложка, ложкой суп едят. Кошка – это кошка, у кошки семь котят. Тряпка – это тряпка, тряпкой вытру пол. Шапка – это шапка, оделся и пошел* (И.Токмакова).

[8] О ситуации множественности референтов у одного собственного имени см. также в работе [Иомдин, Бердичевский 2006: 196–197].

026-Iomdin_B:_Layout 1 18.05.2008 14:51 Page 177

Список литературы

1. Бабаева Е.Э. Формирование семантической структуры слова простой в русском языке // Апресян Ю.Д., Апресян В.Ю., Бабаева Е.Э., Богуславская О.Ю., Иомдин Б.Л., Крылова Т.В., Левонтина И.Б., Санников А.В., Урысон Е.В. Языковая картина мира и системная лексикография. М: Языки славянских культур, 2006. С. 761–844.

2. Богуславская О.Ю. Похожий, схожий, сходный, подобный // Новый объяснительный словарь синонимов русского языка. Второе издание, исправленное и дополненное. Под общим руководством акад. Ю. Д. Апресяна. М.-Вена: Языки славянской культуры; Wiener Slawistischer Almanach. Sonderband 60, 2004. С. 801–812.

3. Иомдин Б.Л., Бердичевский А.С. А кто этот этот? Имена собственные и неопределенная определенность // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.). / Под ред. Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея. М.: Изд-во РГГУ, 2006. С. 196–201.

4. МАС – Словарь русского языка в четырех томах. М.,1981–1984.

5. СОШ – Ожегов С.И., Шведова Н.Ю. Толковый словарь русского языка. М., 1992.

6. СУш – Толковый словарь русского языка / Под ред. Д. Н. Ушакова. М., 1934–1940.

7. Урысон Е.В. Друг, товарищ, приятель // Новый объяснительный словарь синонимов русского языка. Второе издание, исправленное и дополненное. Под общим руководством акад. Ю. Д. Апресяна. М.-Вена: Языки славянской культуры; Wiener Slawistischer Almanach. Sonderband 60, 2004. С. 297–299.

8. Шмелев А.Д. Референциальные механизмы русского языка. Тампере, 1996.

027-Iomdin_L:_Layout 1 18.05.2008 14:52 Page 178

В ГЛУБИНАХ МИКРОСИНТАКСИСА: ОДИН ЛЕКСИЧЕСКИЙ КЛАСС СИНТАКСИЧЕСКИХ ФРАЗЕМ[1]

IN THE DEPTHS OF MICROSYNTAX: A LEXICAL CLASS OF SYNTACTIC IDIOMS

Иомдин Л.Л. *(iomdin@iitp.ru)*
Институт проблем передачи информации РАН

С теоретической и прикладной точек зрения рассматривается класс русских синтаксических фразем, сформированных с участием существительного сила. Каждая такая фразема обладает неповторимым набором свойств. Предлагаются индивидуальные описания нескольких из этих фразем, которые строятся по единой схеме.

1. Вводные замечания

Продолжая серию статей, посвященных синтаксическим фраземам русского языка (см. о них, в частности, [1-6]), в настоящей работе автор развивает и дополняет микросинтаксический подход к описанию таких языковых единиц. Своими основными задачами автор видит, во-первых, по возможности полное описание конкретных синтаксических фразем, т.е. заполнение крупной лексикографической лакуны (для большинства из этих фразем такая работа не проводилась никем и никогда), а, во-вторых, поиск решений, которые позволили бы использовать такие описания в прикладных задачах автоматической обработки текстов (в первую очередь – автоматического синтаксического анализа). Следует сразу же отметить, что образцом лексикографических описаний синтаксических фразем для автора является Новый объяснительный словарь синонимов русского языка Ю.Д. Апресяна и его коллег [7], однако, специфика материала обусловливает определенную модификацию схем таких описаний.

Ниже будут с достаточной степенью подробности рассмотрены следующие фразеологические единицы, образованные с помощью существительного *сила:*

1) предлог ***В СИЛУ*** (*В силу этой теории поведение в одной точке вселенной влияет на поведение в другой точке*);

2) наречие степени ***ОТ СИЛЫ*** (*от силы десять человек*);

3) наречная фразеосхема ***В Х-ОВУЮ СИЛУ*** (*Работает в полную силу, в треть силы, вполсилы*);

4) предикативное наречие ***В СИЛАХ 1*** (*Старик был не в силах быстро ходить*);

5) предикативная наречная фразеосхема ***В СИЛАХ 2*** (*сдержать смех было не в моих силах*).[2]

Нетрудно увидеть, что выше перечислены далеко не все фразеологические единицы, образуемые с участием слова *сила*: в этом легко убедиться, открыв любой толковый словарь русского языка, где отмечается гораздо большее число таких единиц. Данная выборка обусловлена стремлением продемонстрировать, во-первых, разнообразие в поведении даже близких по смыслу фразеологических единиц, а, во-вторых, дать читателю представление о масштабе еще не решенных задач, стоящих перед лексикографами, взявшимися за систематическое описание русского микросинтаксиса.

При работе над данной статьей автор широко использовал материалы Национального корпуса русского языка (www.ruscorpora.ru), а также Корпуса русского языка, имеющегося в секторе теоретической семантики Института русского языка РАН.

Из-за ограниченного объема публикации фразеологические единицы будут в основном характеризоваться неформально: полное описание этих единиц, включая формальные толкования, потребовало бы существенно большего пространства.

---

[1] Данное исследование выполнено в рамках трех проектов, подержанных Российским фондом фундаментальных исследований (гранты № 06-04-00090, 07-06-00339 и 08-06-00373). Автор пользуется случаем, чтобы выразить Фонду свою признательность.

[2] Строго говоря, не все эти единицы можно с полным правом отнести к синтаксическим фраземам в смысле [8] или [9]: единицы *в силу* и *от силы* являются рядовыми составными предлогами и наречиями и не отличаются значительной синтаксической спецификой от других таких единиц. Мы рассматриваем их здесь потому, что представляет особую проблему идентификация этих единиц в тексте вследствие порождаемой этими единицами синтаксической омонимии и их отграничение от других синтаксических фразем.

027-Iomdin_L:_Layout 1 18.05.2008 14:52 Page 179

2. Составной предлог В СИЛУ

2.1. Синтаксис. Заметим прежде всего, что составной предлог *в силу*, образованный из первообразного предлога *в* и существительного *сила* в вин. падеже ед. числа, выступает в русском языке как полноценный, почти первообразный предлог, управляющий родительным падежом. В отличие от многих других предложных выражений, образованных из прототипического предлога и существительного, таких как *в виде, в лице, в качестве, в случае*, *в отношении, за исключением, за счет, на основании, по поводу, с помощью, в связи с* и др. данный предлог не может разрываться даже местоименными определениями к существительному и тем самым никогда не переходит в наречное речение. Ср. *Он работает в качестве инженера с прошлого года – Он работает в этом качестве с прошлого года*; *На основании какого документа мне отказано? – На каком основании мне отказано?* Для *в силу* это невозможно: *В силу этой теоремы данное утверждение верно,* но не **В ее силу данное утверждение верно.*[3] С этой точки зрения предлог *в силу* относится к тому же (меньшему, чем первый) классу русских составных предлогов, что и *в отличие от, в противовес, в угоду, в ходе, по сравнению с.*

Далее, первообразность предлога *в силу* выражается в том, что личные местоимения третьего лица, подчиняющиеся этому предлогу, практически обязательно приобретают начальное *н-,* как в случае первообразных предлогов, ср.

(1) *Иногда в силу теории вероятности выпадают подряд одинаковые карты – Иногда в силу нее выпадают подряд одинаковые карты*,

но не **в силу ее….*

Тем не менее у *в силу* существуют синтаксические особенности, не позволяющие безоговорочно квалифицировать его как первообразный предлог, а именно: в сочетаниях с некоторыми парными союзами (*либо… либо, то ли… то ли, не то… не то, как.. так и*) или частицами *ли и же* эти последние могут размещаться между данным предлогом и зависящим от него существительным; ср. *В силу как обстоятельств, так и личных пристрастий Набоков стремится осознавать себя как абсолютно изолированную единицу, существующую вне культуры как таковой*; *Не в силу ли этого своего хобби он создал равное количество романов по-русски и по-английски?* (из статей о В.Набокове); *Гуманитарные ценности, в силу ли своего облагораживающего влияния на человечество, в силу ли объединения его вокруг общих ценностей, в силу ли создания идеалов эквивалентны по ценности естественнонаучным; В силу же полной обратимости времени аксиома III отсутствует в механике Ньютона.* Ничего подобного настоящие первообразные предлоги не допускают: **Не из-за ли своего хобби…* (надо: *не из-за своего ли хобби…).*

2.2. Семантика и сочетаемость. Предлог *в силу*, без сомнения, не композиционален относительно своих составных частей. В смысл этого предлога не входит ни смысл первообразного «направительного», управляющего винительным падежом предлога *в* (во всяком случае, в его основном значении), ни смысл существительного *сила* в каком бы то ни было из его лексических значений.

Примерное толкование этого предлога могло бы выглядеть следующим образом:

(2) *P в силу X-а* ≈ ‘Имеет место факт, свойство, событие или ситуация X (пресуппозиция); факт, свойство, событие или ситуация P объясняется тем, что имеет место X (ассерция)’.

Следует отметить, что традиционные толковые словари русского языка относят *в силу* к причинным предлогам (например, в словаре С.И.Ожегова *в силу* толкуется как ‘по причине чего-н., из-за чего-н.’). Однако это неверно или, во всяком случае, неточно: в действительности данный предлог употребляется тогда, когда причинная связь между X и P если и имеется, то носит достаточно сложный и опосредованный характер; для экспликации этой связи требуются логические рассуждения; неформально говоря, употребляя предлог *в силу,* говорящий предлагает слушающему произвести эти логические рассуждения. Заметим, например, что замена предлога *в силу* в предложении (1) на основной русский причинный предлог *из-за* приведет к существенному изменению смысла: выпадение одинаковых карт происходит не потому, что имеет место теория вероятностей (ее могло бы вообще не существовать, что не отразилось бы на поведении карт); говорящий же в (1) объясняет выпадение одинаковых карт соответствием этого события теории вероятностей.

В соответствии с толкованием (2) вряд ли уместны высказывания типа *Я не успел в театр в силу опоздания поезда*: в подобной ситуации причинно-следственная связь между двумя событиями слишком очевидна, чтобы для ее описания прибегать к тяжелой артиллерии книжного предлога *в силу.*

Существенно, с другой стороны, что, подобно причинным предлогам, предлог *в силу* вводит неравноправные факты: факт X признается первичным, независимым, а факт P – вторичным. Поэтому

[3] Заметим для полноты картины, что кристаллизация этого предлога произошла лишь недавно. Еще в XIX веке ситуация была другой и можно было встретить выражения типа *Я предложил моему наставнику такой вопрос: в какую силу принимаются им все эти приношения* (И.С.Никитин, 1860). Более того, даже в наши дни возможна языковая игра с данным предлогом (пусть и не в лучшем вкусе): *В эту же силу к множественному принадлежит профанный нуль* (из Интернета).

027-Iomdin_L:_Layout 1 18.05.2008 14:52 Page 180

выражения типа *в соответствии друг с другом* допустимы, а выражения **по причине друг друга* и **в силу друг друга* – нет. В предложении *Они [творческие индивиды] либо делают подлости в силу глупости, либо глупости в силу подлости* (А. Зиновьев, «Зияющие высоты») описываются две разные ситуации, в которых факты X и P меняются местами, но не одна ситуация взаимообусловленности X и P.

В полном соответствии с толкованием, предлог *в силу* подчиняет реализующие валентность X существительные со значением фактов, свойств, событий и ситуаций (но, например, не предметов и лиц: **в силу Америки,* * *в силу Ивана*) и подчиняется реализующим (пассивную) валентность P глаголам и существительным с теми же значениями.

2.3. Идентификация предлога *в силу* при автоматической обработке текстов. Как и в других ситуациях, когда неоднословная фразеологическая единица языка омонимична свободным словосочетаниям и/или другим фразеологическим единицам, надежное отождествление ее в тексте при автоматическом анализе представляет собой сложнейшую и трудно разрешимую проблему. Не говоря уже о случаях реальной неоднозначности предложения (ср. *Он верит в силу привычки* ‘предмет его веры есть сила привычки’, vs. ‘его вера [например, в Бога] объясняется привычкой’), отграничение в тексте предлога *в силу* от последовательности двух слов *в* и *силу* не может быть описано с помощью сколько-нибудь надежных правил.

Данное утверждение подтверждается следующим небольшим экспериментом, проведенным автором. Автор проверил встречаемость в текстах некоторых сочетаний цепочки *в силу* со словами, которые, по его эмпирическим ощущениям, максимально благоприятствовали интерпретации этой цепочки как предлога: 1) *в силу теоремы <леммы, аксиомы, гипотезы, утверждения>*, 2) *в силу закона, теории, разума, ума, слова.* Выяснилось, что если в первой группе словосочетаний *в силу* действительно практически всегда следует интерпретировать как предлог, то во второй группе ситуация существенно меняется: *в силу теории* содержит предлог *в силу* в 90% случаев (но ср., например, непредложное функционирование во фразах *Он верил в силу теории* или *Тут вступают в силу теории эволюции*), *в силу ума* содержит предлог в 70% случаев, *в силу закона* – в 20% случаев (здесь возмущающим элементом является устойчивое сочетание *вступление в силу*), а *в силу разума* – лишь в 8% случаев (ср. *Индивидуальность в силу разума обладает способностью к реструктуризации*). Добавим, что стопроцентное попадание выражения *в силу* в предлог наблюдается в цепочке *в силу чего* (это начало изъяснительного придаточного) и почти стопроцентное – в цепочке *в силу чьего* (отмечено одно исключение: *Вы только вслушайтесь в силу чьего-то чувства к неведомой нам девушке*).

Отсюда можно сделать следующий вывод. Частично информацию о совместной встречаемости предлога с каким-либо словом можно учесть в статистических моделях автоматического анализа текста. Если представить себе, однако, объем предшествующей такому учету эмпирической работы и помножить его на количество фразеологических единиц, заслуживающих такой же эмпирической работы, станет ясно, что практическое решение подобного класса задач в сколько-нибудь полном объеме возможно лишь после создания словаря микросинтаксических конструкций.

Сказанное в данном разделе в полной мере относится ко всем другим фразеологическим единицам, излагаемым выше. Из соображений экономии места мы не будем приводить подобные рассуждения для этих единиц.

Заключим эту главку следующей курьезной информацией. Еще в тридцатых-сороковых годах XX века русская орфография допускала слитное написание данного предлога – *всилу,* что отмечено в толковом словаре Д.Н.Ушакова. Если бы творцы орфографической нормы учитывали интересы создателей систем автоматической обработки текстов, работа последних могла бы стать существенно легче.

3. Наречие степени ОТ СИЛЫ

3.1. Синтаксис. Данное наречие, сформированное предлогом *от* и существительным *сила* в родительном падеже единственного числа, выступает в качестве препозитивного или постпозитивного модификатора при именных группах и, изредка, глаголах, ср.

(3) *Он получит от силы 1000 рублей <1000 рублей от силы>*;

(4) *За это его от силы пожурят, но, конечно, не уволят.*

От силы не имеет существенных синтаксических особенностей, отличающих его от других степенных наречий, особенно таких, которые сочетаются с числовыми выражениями (*максимум, максимально, минимум* и др.; ср., с одной стороны (3)-(4), а с другой высказывания *Он получит максимум 1000 рублей, За это его максимум поставят в угол*). Мы сосредоточимся поэтому на семантических свойствах этого наречия.

3.2. Семантика и сочетаемость. Как и предлог *в силу*, наше наречие не композиционально относительно своих составных частей. Автор не видит в семантике этого наречии даже следа значений формирующих его лексических единиц – предлога *от* или существительного *сила.*

027-Iomdin_L:_Layout 1 18.05.2008 14:52 Page 181

Приблизительное толкование этого наречия выглядит так:

(5) *От силы P* ‘максимальная величина некоторого параметра Q составляет или может составить P, говорящий считает, что эта величина мала’.

Обе валентности наречия *от силы* – пассивные. В предложении (3) валентность P заполнена именной группой *1000 рублей*, а валентность Q – глаголом *получит*. В предложении (4) валентность P заполнена глаголом *пожурят*, а валентность Q эксплицитно не выражена, хотя примерный ее смысл ясен любому носителю языка – это ‘наказание’.

Имплицитный характер валентности Q сближает наше наречие с подробно рассматриваемыми в [11] кванторными словами типа *только*. Семантическая экспликация предложения (3) посредством нашего толкования (5) могла бы выглядеть так: ‘Из всех видов наказаний самое строгое наказание, которое он за это понесет, – его пожурят; говорящий считает, что это недостаточное наказание’.

Пресуппозитивный компонент толкования (3) ‘говорящий считает, что эта величина мала’ представляется весьма важным: он отличает выражение *от силы* от его ближайших синонимов *максимум* и *максимально*. Ср., например, предложения

(6) *На этом турнире любая команда может набрать максимум 10 очков*,

где говорящий предлагает объективный арифметический подсчет результатов (при этом оценка говорящим величины в 10 очков может быть и положительной, и отрицательной, и вообще отсутствовать), и

(7) *На этом турнире любая команда может набрать от силы 10 очков*,

где говорящий со всей определенностью характеризует эту величину как недостаточную.

Добавим к сказанному, что составное наречие *от силы* разделяет со своими синонимами и аналогами (*максимум, минимум, как минимум* и пр.) способность задавать разные ориентации иерархической шкалы некоторых параметров (см. об этом [12]): от низшего точки к высшей или наоборот. Предложение *Она займет от силы третье место* означает, что он может занять третье, четвертое и последующие места, но не первое и не второе, а предложение *Каждый из нас помнит имена своих родственников от силы до четвертого-пятого поколения* означает, что не помнятся имена родственников шестого и более далеких поколений.

В заключение отметим, что наречие *от силы*, как и другие разговорные лексические единицы со значением количественной оценки, охотно сочетаются с аппроксимативно количественными конструкциями: (*Придет от силы человек восемь*; *Я заработал от силы тысяч десять*. Весьма любопытно при этом, что замена в таких высказываниях аппроксимативной конструкции на выражение, содержащее эксплицитное указание на приблизительность, приводит к неправильности высказывания: **Придет от силы приблизительно восемь человек, *Я заработал от силы примерно десять тысяч*. По ощущениям автора, эта неправильность носит не чисто сочетаемостный или грамматический, а семантический характер, однако для ее объяснения требуется дополнительное исследование. Укажем пока на то обстоятельство, что сам по себе пресуппозитивный компонент смысла *от силы* ‘эта величина мала’ не может быть причиной указанной неправильности: выражения типа *Я заработал только <всего> примерно десять тысяч* вполне правильны, хотя *только* и *всего* тоже содержат указанный компонент.

4. Наречная конструкция В Х-ОВУЮ СИЛУ

4.1. Синтаксис. Наречная конструкция, или, если воспользоваться термином Д.Н. Шмелева [13], фразеосхема *в Х-овую силу* обладает той важной особенностью синтаксической фраземы, что содержит как постоянную, так и обязательную переменную часть. Постоянная часть образована управляющим винительным падежом предлогом *в* и существительным *сила* (на этот раз в одном из своих основных лексических значений), а переменная часть представлена словами лексического класса весьма причудливого состава, характеризующими степень применения этой силы. В этот лексический класс входят прилагательные *полный, неполный, половинный;* количественные существительные *половина, половинка, треть, четверть, часть, третья, десятая* и т.п (в значении *третья часть, десятая часть*), а также числительные – *пол, полторы, две, три, девять* и т.п. Синтаксические связи внутри данной фраземы оказываются различными в зависимости от того, как реализована переменная часть – если это прилагательное или числительное, то оно подчиняется существительному *сила,* а если это количественное существительное, то оно подчиняется предлогу *в* и подчиняет себе существительное *сила*. Если переменная часть реализуется существительным *часть,* то оно почти всегда сопровождается порядковым прилагательным: *в пятую часть силы* и т.п.

Разумеется, частоты встречаемости в текстах конкретных реализаций переменной части нашей синтаксической фраземы несопоставимы друг с другом – прилагательное *полный* используется регулярно, а все остальные варианты появляются в единичных случаях и чаще всего в качестве метафоры или в порядке языковой игры. Показательный пример: в критической статье о романе В.Орлова, озаглавленной «Альтист Данилов, игра

027-Iomdin_L:_Layout 1 18.05.2008 14:52 Page 182

не в полную силу», где употреблен обычный вариант нашей фраземы, встречается и синонимичное выражение *Он играл в неполную силу.* Ср. также другие примеры, где налицо явная языковая игра: *В лагере каждый тянет вполсилы*[4] *или в полторы силы. Дружно в лагере тянуть не умеют* (В.Шаламов); *Российским футболистам будет сложно справиться с реактивными «камикадзе», играющими не просто в полную силу — в две силы*; *Сверхдержава воевала даже не в треть, а в десятую силы*; *Семнадцать лет. По улицам кружить, / Читать Ахматову и Гумилева, / Дышать вполсилы, в четверть силы жить* (Б. Верникова) и т.д.

4.2. Семантика. Единица *в Х-овую силу в* целом ведет себя как наречие степени и служит модификатором предикатного слова (прототипически – глагола или отглагольного существительного).

Примерное толкование синтаксической фраземы *в Х-овую силу* выглядит так:

(8) *Р в Х-овую силу* ‘действие или деятельность Р происходит так, что субъект Р использует при совершении Р часть своих возможностей, равную Х’.

Существенной семантической особенностью нашей фраземы является тот факт, что формально количественная[5] оценка прилагаемой силы фактически всегда используется как метафора: если говорится, например, что один футболист играл матч вполсилы, другой в треть силы, а третий - в четверть силы, то никакого реального подсчета при этом не производится, утверждается лишь, что футболисты играли ниже своих возможностей. Тем самым можно утверждать, что – во всяком случае, с точки зрения узуса – выражения типа *в треть силы* и *в четверть силы* синонимичны друг другу, как и выражения типа *в две силы и в три силы.* Разумеется, такая ситуация наблюдается и в других фразеологических единицах и просто в речевых штампах: выражения *Я сто раз тебе повторял, что так делать нельзя* и *я тысячу раз тебе повторял, что так делать нельзя* значат одно и то же.

Автору неясно пока, следует ли относить к рассматриваемой синтаксической фраземе выражения, в которых переменная часть представлена существительными типа *разряд, гроссмейстер* и т.п.; ср. *Сюсаку к тому времени уже играл в силу 7-го дана; Зрелищным оказался турнир «молодых», большинство из которых играют в силу крепкого I разряда; Если Чигорин играл в силу международного мастера, то Ласкер образца 1905-го играл в силу очень приличного гроссмейстера,* а также *Каждый работал в силу своих возможностей.* С одной стороны, такие выражения характеризуют уровень приложения сил, однако скорее они характеризуют его не с количественной стороны, а с качественной и вводят идею сравнения с эталоном.

5. Предикативное наречие В СИЛАХ 1 и предикативная наречная фразеосхема В СИЛАХ 2

5.1. Синтаксис. Насколько известно автору, никакие словари и грамматики русского языка не различают двух фразеологических единиц *в силах.* Между тем, как мы сейчас увидим, с синтаксической точки зрения эти две единицы резко контрастируют друг с другом.

Синтаксическая фраземa *в силах 1*, образованная предлогом *в*, управляющим предложным падежом, и существительным *сила* в предложном падеже множественного числа – это предикативное наречие, выполняющее в предложении роль сказуемого (точнее, именной части составного сказуемого, глагольная часть которого представляет собой связку). Подлежащее при таком сказуемом стоит в именительном падеже; кроме того, эта синтаксическая фраземa имеет дополнение, выражаемое инфинитивом, ср. *Иван в силах сделать это; Я оказался не в силах побороть усталость*; *Больной к ночи уже был не в силах поднимать рук и только смотрел пред собой, не изменяя внимательно сосредоточенного выражения взгляда* (Л.Н. Толстой); *Я, не будучи в силах преодолеть себя, иду к рыжему дому и звоню к дворнику* (А.П. Чехов).

С синтаксической точки зрения фраземa *в силах 1* ведет себя так же, как синтаксическая фраземa *в состоянии* (ср. *Я в состоянии сделать это*), и весьма близко к прилагательному *способен* (*Я способен сделать это*).

Синтаксическая фраземa *в силах 2* ведет себя существенно иначе. Хотя эта фраземa тоже является предикативным наречием и выполняет роль именной части составного сказуемого, она отличается от *в силах 1,* во-первых, тем, что, подобно рассмотренному в п. 4 наречию *в силу,* имеет обязательную переменную часть и тем самым является фразеосхемой. Эта переменная часть выражается существительным в родительном падеже или прилагательным (чаще всего местоименным, но не только), которое вводит субъекта данной фраземы: *в силах Ивана, не в моих силах*. Во вторых, подлежащим при сказуемом *в силах 2* прототипически выступает не именная группа, а инфинитив; ср. *Не в моих силах писать по-французски* (В. Некрасов). Тем самым фраземе *не в силах 2* при ее описании в терминах синтаксического компонента модели «Смысл-Текст» должен быть приписан особый

[4] Тот факт, что по прихоти русской орфографии данный вариант фразеосхемы пишется слитно, мы оставляем без внимания.

[5] И иногда весьма точная оценка, ср. *Говоря со мной об одном человеке, он сказал, что ум у него в девять сил из расчета на морскую свинку*; *Бил он в сотую часть силы, но у Аникиной перехватило дыхание* (Русский национальный корпус).

027-Iomdin_L:_Layout 1 18.05.2008 14:52 Page 183

синтаксический признак ПРЕДИНФ (см. о нем в [14]).

Нетрудно заметить, что с синтаксической точки зрения фраземы *в силах 1* и *в силах 2* практически являются конверсивами.

К сказанному необходимо сделать еще три добавления.

Первое: фразеосхема *в силах 2,* помимо переменной части, может еще сопровождаться эпитетом к слову *силах,* что представляется достаточно необычным свойством; ср. *Сделать это не в моих скромных <слабых> силах.*

Второе. Мы видели, что и при *в силах 1*, и при *в силах 2* появляется инфинитив – в первом случае он реализует собственное управление фраземы, а во вторых – несобственное управление через связку. Характерно при этом, что в первом случае инфинитив не альтернирует даже, как это обычно бывает, с местоименным словом; ср. *Он любит решать задачи – Он любит это – Что он любит,* но *Он в силах решить задачи – *Он в силах это – *Что он в силах.* В случае же *в силах 2* инфинитив, выполняющий роль подлежащего, чередуется не только с местоимением, но и с другим словом широкой семантики; ср. *Сделайте все, что в человеческих силах. В человеческих силах немалое, а потому трое способнейших, запершись в кабинете, начали там систематический обыск* (Карел Чапек, пер. Т. Аксель и О. Молочковского).

Третье. Как было только что показано, две достаточно близких по смыслу фраземы, мотивированные одним и тем же словом, обладают весьма различными синтаксическими свойствами. Если ввести в рассмотрение еще две фраземы с близкими значениями – *под силу* и *по силам* (увы, в данной публикации для них нет места), то обнаружится, что их свойства отличаются как от *в силах 1, так и в силах 2.* Главное отличие состоит в том, что субъект этих фразем выражается дательным падежом, превращая конструкции в безличные: *Мне не под силу <не по силам> выполнить такую работу за месяц.* Данный факт – еще одна демонстрация гигантского объема работы, который предстоит выполнить микросинтаксистам.

5.2. Семантика. В данной главке мы ограничимся лишь неформальным указанием на одно важное различие в семантике фразем *в силах 1* и *в силах 2*, оставляя на будущее их подробное описание, включая толкования.

В силах 1 свободно употребляется для характеристики любых способностей субъекта – как физических, так и умственных, психических, моральных и т.д: ср. *Я (не) в силах поднять этот чемодан – Врач был не в силах помочь безнадежному больному.* Эти способности воспринимаются как внутренне присущие субъекту (в первую очередь, разумеется, человеку).

Что касается *в силах 2,* то эта фразема характеризует в первую очередь моральные способности человека, причем представляет их как не зависящие от его воли, как бы данные ему свыше. Поэтому вряд ли уместно сказать что-то вроде *Поднять этот чемодан в его силах* или *Поднять этот чемодан не в его силах*: в первом случае говорящий представил бы достаточно рядовое физическое здоровье субъекта как данное ему высшей силой, а во втором случае, соответственно, он представил бы недостаток физической силы у субъекта как результат воздействия некоторого высшего существа. И то, и другое в бытовой ситуации не оправданно.

Список литературы

1. Л.Л.Иомдин. Большие проблемы малого синтаксиса // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям Диалог’2003. Протвино, 2003. С. 216-222.

2. Leonid Iomdin. A Hypothesis of Two Syntactic Starts. // East-West Encounter: Second International Conference on Meaning - Text Theory. Edited by Ju.D. Apresjan and L.L. Iomdin. 165-175. Slavic Culture Languages Publishing House. Moscow.

3. Л.Л.Иомдин. Многозначные синтаксические фраземы: между лексикой и синтаксисом // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции «Диалог-2006». Москва: Изд-во РГГУ, 2006. С. 202-206.

4. Л.Л.Иомдин. Новые наблюдения над синтаксисом русских фразем // Obecność. Red. Bożena Chodźko, Elżbieta Feliksiak, Marek Olesiewicz. Białystok: Uniwersytet w Białymstoku, 2006. S. 247-281.

5. Leonid Iomdin. Russian Idioms Formed with Interrogative Pronouns and their Syntactic Properties // Meaning – Text Theory 2007. Proceedings of the 3rd International Conference on Meaning – Text Theory. Wiener Slawistischer Almanach. Sonderband 69. München – Wien, 2007. ISSN 0258-6835. ISBN 978-3-87690-xxx-x. S. 179-189.

6. Л.Л. Иомдин. Русские конструкции малого синтаксиса, образованные вопросительными местоимениями. // Мир русского слова и русское слово в мире. Материалы XI Конгресса Международной ассоциации преподавателей русского языка и литературы. Heron Press. Sofia 2007. ISBN 978-954-580-213-3. С. 117-126.

027-Iomdin_L:_Layout 1 18.05.2008 14:52 Page 184

7. Ю.Д. Апресян и др. Новый объяснительный словарь синонимов русского языка. Второе издание, исправленное и дополненное. Под общим руководством акад. Ю. Д. Апресяна. Москва-Вена: 2004. Языки славянской культуры; Wiener Slawistischer Almanach. Sonderband 60.

8. И.М. Богуславский, Л.Л. Иомдин. Безусловные обороты и фраземы в толково-комбинаторном словаре // Актуальные вопросы практической реализации систем автоматического перевода. Ч. 2 М.: Изд-во МГУ, 1982. С. 210-222.

9. Geoffrey Nunberg, Ivan A. Sag, Thomas Wasow. Idioms. In: Language, 1994, Vol. 70, 491-538.

10. Igor Mel'čuk. Phrasemes in language and phraseology in linguistics. In. Idioms: Structural and Psychological Perspectives, chapter 8. M. Everaert, E-J. van der Linden, A. Schenk, and R. Schreuder, editors. Lawrence Erlbaum Associates.1995.

11. И.М. Богуславский. Сфера действия лексических единиц. М.: Школа «Языки русской культуры». 1996.

12. Б.Л. Иомдин, Л.Л. Иомдин. Семантика экстремумов // Четвертая типологическая школа (Четвертая международная школа по лингвистической типологии и антропологии. Ереван, 21-28 сентября 2005 г.). Материалы лекций и семинаров. Ред. В.И. Подлесская (отв.), А.В.Архипов, Ю.А.Ландер. М.: Издательский центр РГГУ, 2005. С. 167-171.

13. Д.Н. Шмелев. О "связанных" синтаксических конструкциях в русском языке. // Вопросы языкознания. 1960, № 5. С. 47-60.

14. Л.Л. Иомдин, И.А. Мельчук, Н.В. Перцов. Фрагмент модели русского поверхностного синтаксиса. I. Предикативные синтагмы // Научно-техническая информация. Серия. 2. 1975. № 7. С. 30-43.

028-Kibrick_AA:_Layout 1 18.05.2008 15:13 Page 185

ПРОСОДИЧЕСКИЙ ПОРТРЕТ ГОВОРЯЩЕГО КАК ИНСТРУМЕНТ ТРАНСКРИБИРОВАНИЯ УСТНОГО ДИСКУРСА[1]

SPEAKER'S PROSODIC PORTRAIT AS A TOOL OF SPOKEN DISCOURSE TRANSCRIPTION

Кибрик А.А. *(kibrik@comtv.ru)*
Институт языкознания РАН

В докладе предлагается методологический аппарат, повышающий качество транскрибирования устного дискурса при создании корпусов звучащей речи. Просодические звукотипы, лежащие в основе сегментации дискурса и выражения фазовых значений, идентифицируются при помощи просодических портретов индивидуальных говорящих.

1. Вводные замечания

В настоящее время появляется все больше корпусов устной речи на разных языках. Это важный и позитивный процесс – он связан с признанием того факта, что устная форма языка как минимум не менее важна, чем письменная. При этом создание устных корпусов сопряжено с большими трудностями, ведь главный компонент устного корпуса – это даже не собственно звук, а *транскрипт* звука. Транскрибирование устного дискурса, то есть преобразование звучащего дискурса в транскрипт, включает в себя множество решений, которые должны быть систематическими, последовательными и универсальными. Если транскрибирование осуществляется ad hoc, без надлежащей теоретической базы, то ценность получаемого продукта невелика, этот продукт не отражает сущностных свойств исходного объекта, то есть устного дискурса.

Данный доклад основан на опыте разработки транскрипции устного дискурса, связанном с подготовкой устного русского корпуса «Рассказы о сновидениях», см. Кибрик и Подлесская 2003. Одна из задач этого проекта состоит в том, чтобы отразить специфику организации устного дискурса при помощи дискурсивной транскрипции.

2. Сегментация и фаза

К числу важнейших компонентов дискурсивной транскрипции относятся сегментация дискурса и выражение фазовых значений (см. Кибрик и Подлесская 2006). Дискурс продуцируется говорящим не в виде плавного потока, а в виде квантов или сегментов – элементарных дискурсивных единиц (ЭДЕ). ЭДЕ идентифицируются при помощи комплекса просодических параметров, включая паузацию, темп, громкость, наличие единого тонального контура и главного акцента. В литературе по просодической сегментации устной речи чаще используются другие термины – фраза, интонационная фраза, интонационная группа, ритмическая группа, интонационная единица (см., например, Светозарова и др. 1988; Chafe 1994: 57; Хитина 2004; Кривнова 2007: раздел 2.3).

Как правило, каждая ЭДЕ маркирована с точки зрения *фазы* (термин из работы Кодзасов 2002), то есть по признаку конечности/неконечности. Можно различать разные иерархические уровни фазовых значений. Так, иллокутивная фаза связана с конечностью/неконечностью в коммуникативном взаимодействии: общий вопрос является неконечной иллокуцией, сообщение – конечной. В рамках одной иллокуции ЭДЕ также квалифицируются как конечные или неконечные; этот тип фазы можно назвать внутренним. Основным средством маркирования фазы в устном дискурсе является движение тона в главном акценте (см. Кодзасов 2002). Говоря наиболее обобщенно, конечная иллокутивная фаза маркируется нисходящим (падающим) тоном, неконечная иллокутивная – восходящим, а неконечные внутренние ЭДЕ зеркально адаптируются к конечной.

В дискурсивной транскрипции сегментация на ЭДЕ обычно передается делением на строки, а фаза – при помощи пунктуационных знаков в конце строк. Пунктуационные знаки в этом случае используются не просто как аналоги пунктуации, принятой в письменной форме языка, а как строгие обозначения, расставляемые на основе дискурсивной семантики и просодических фактов. Именно такая система обозначений принята в транскрипции, разрабатываемой в проекте «Рассказы о сновидениях», см. Кибрик и Подлесская 2003, 2009.

[1] Данное исследование выполнено при поддержке гранта РГНФ 08-04-00165а.

028-Kibrick_AA:_Layout 1 18.05.2008 15:13 Page 186

И для сегментации, и для разметки фазы большое значение имеют частотные характеристики, то есть характеристики основного тона голоса говорящего. Приведу два примера. Во-первых, при решении вопроса о границе ЭДЕ используется понятие нейтрального тонального уровня, с которого говорящий начинает обычную ЭДЕ; выход на такой нейтральный уровень иногда называют «ресет». Наиболее типичный тональный паттерн состоит в том, что после начала с нейтрального уровня происходит подъем тона, а к концу деклинация. Во-вторых, при решении вопроса о конечности/неконечности данной ЭДЕ в рамках иллокуции используется представление о самом низком уровне тона, характерном для данного говорящего. Дело в том, что в речи большинства носителей русского языка имеются два семиотически противопоставленные типа падающего тонального акцента:

- конечный – так называемая интонация точки: падение в нижний уровень голоса, комфортный для данного говорящего
- неконечный – так называемая интонация запятой с падением: падение в уровень, на 2-4 полутона выше нижнего

3. Относительность просодических характеристик

Таким образом, оба рассмотренные просодические явления – ресет и тип падения – являются не абсолютными, а относительными, зависят от характеристик голоса индивидуального говорящего. Если транскрайбер (то есть эксперт, выполняющий транскрибирование устного дискурса) не располагает информацией о голосе данного говорящего, он не может принять решение о том, является ли уровень тона в начале данной ЭДЕ ресетом, и о том, является ли данное падение конечным или неконечным.

Это напоминает общеизвестную ситуацию, связанную с различием между фонетикой и фонологией. Когда носитель языка сталкивается с конкретным звуковым сегментом, в огромном большинстве случаев он в состоянии идентифицировать эту конкретную инстанцию как экземпляр фонемы. Как именно эта идентификация происходит – предмет научного исследования. Одни носители шепелявят, другие грассируют, третьи склонны к фарингализации и т.д. Реализация фонем чрезвычайно вариативна по говорящим и – в рамках одного говорящего – по различным фонетическим контекстам, по функциональным стилям, степени формальности речи и многим другим параметрами. Тем не менее, нормальный носитель языка справляется с этой задачей без больших усилий. Лингвисту, даже имеющему в своем распоряжении арсенал акустической аппаратуры, не так легко смоделировать этот процесс распознавания. Однако ему помогает его собственная интуиция как носителя и достаточно надежные знания о составе фонем языка и о возмущающих факторах.

Те же проблемы имеются в области несегментных аспектов звука – просодии. Носитель языка, конечно, легко распознает все значимые просодические феномены и понимает их семантику. А положение лингвиста здесь заметно сложнее. Во-первых, сами просодические звукотипы известны менее надежно, чем в сегментной сфере. Во-вторых, отсутствует очевидный метаязык, который описывал бы особенности конкретных говорящих. Данный доклад связан с последним обстоятельством. Прежде чем принимать решения о сегментации и расстановке пунктуационных знаков в транскрипте дискурса данного говорящего, необходимо вначале изучить просодическую систему этого человека. Описание наиболее важных элементов этой системы можно назвать *просодическим портретом*. Можно сказать, что рассматриваемая здесь задача – необходимость увидеть просодическую фонологию за просодической фонетикой.

4. Компоненты просодического портрета

С точки зрения принятия решений о сегментации и об отражении фазовых значений при работе над корпусом «Рассказы о сновидениях» наиболее важными оказались следующие элементы просодического портрета говорящего:

- тональный диапазон
- типичный тон начал (ресетов)
- целевой уровень конечных падений
- целевой уровень неконечных падений
- целевой уровень подъемов при запятых
- уровень заударного падения при запятых
- уровень подъема при многоточиях

Тональный диапазон характеризует голос говорящего в наиболее грубой форме. Этот диапазон показывает минимальные и максимальные значения частот, которые голос говорящего принимает на протяжении дискурса, имеющегося в распоряжении исследователя.

028-Kibrick_AA:_Layout 1 18.05.2008 15:13 Page 187

Типичный тон начал (ресетов) – это нейтральный уровень, с которого говорящему удобно начинать каждую очередную ЭДЕ. Этот тон представляет собой не единичное значение, а определенную полосу. Иначе говоря, он варьирует. Однако это варьирование жестко ограничено. Разница между максимальным и минимальным значениями находится в пределах нескольких полутонов.

Целевой уровень финальных падений – это еще один базовый для говорящего уровень частоты, в который стремится упасть его голос в конце иллокутивного акта сообщения. По мнению О.Ф. Кривновой (личное сообщение), целевой уровень финальных падений совпадает с абсолютным нижним уровнем голоса данного говорящего. Целевой уровень финальных падений также представляет собой некоторую полосу.

В отличие от финальных падений, нефинальные падения происходят в уровень частоты, чуть более высокий. Полоса нефинальных падений иногда четко отличается от полосы финальных падений. Иногда, правда, эти полосы могут пересекаться, подобно тому как реализации некоторых фонем могут подвергаться нейтрализации.

При канонической интонации запятой подъем также происходит в некоторую фиксированную полосу. Размер этой полосы может быть большим: некоторые говорящие временами срываются на фальцет, и тогда интонация запятой может реализовываться очень высоким подъемом тона.

В русской речи (и этим она отличается от многих других европейских языков) при интонации запятой за семиотически значимым подъемом следует автоматическое падение тона. Это падение представлено тогда, когда несущий акцент не приходится на последний слог ЭДЕ; в последнем случае автоматическое падение, как правило, отсутствует. В тех случаях, когда падение есть, представляет интерес его целевая полоса.

Наряду с интонацией точки и запятой, очень часто в корпусе встречается «интонация многоточия». В этом случае чаще всего происходит подъем в уровень, значимо более низкий, чем при канонической интонации запятой. Передаваемая семантика может быть наиболее обобщенно описана как неопределенность фазовой характеризации: говорящий не может определить, является ли ЭДЕ конечной или нет. Значимость этого просодического феномена для исследуемого нами жанра дискурса столь велика, что его также необходимо включить в просодический портрет. При этом опять же оценивается целевая полоса подъема тона при наиболее типичных реализациях многоточия.

Кроме того, в просодический портрет включается пункт «прочие просодические характеристики», где фиксируются различные особенности, которые потенциально могут оказаться важными для того или иного аспекта транскрипции.

5. Примеры просодических портретов

Рассмотрим несколько примеров просодических портретов конкретных рассказчиков. Отметим, что все числовые значения приводятся в герцах (Гц).

тональный диапазон	типичный тон начал (ресетов)	целевой уровень финальных падений	целевой уровень нефинальных падений	целевой уровень подъемов при запятых	уровень заударного падения при запятых	уровень подъема при многоточиях	другие характеристики
160-360	200-220	160-180	190-260	290-360	230-300	210	фрикативное [г]

Таблица 1. Просодический портрет рассказчицы рассказа Z52 (МК ж 16 лет)[2]

Как можно видеть, у данной рассказчицы полоса начал (ресетов) очень узка – она составляет всего лишь 20 Гц. Целевые уровни финальных и нефинальных падений различаются достаточно четко. То же касается различия между подъемом при интонации запятой и подъемом при интонации многоточия. Уровень заударного падения при запятых близок к целевому уровню нефинальных падений.

Эти выводы относительно данной рассказчицы имеют лишь ограниченную надежность, поскольку в корпусе имеется только один принадлежащий ей рассказ. Рассмотрим теперь просодические портреты, основанные на шести рассказах одного и того же рассказчика.

[2] Номера рассказов, в соответствии с нумерацией в текущей версии корпуса, состоят из литеры (например, Z) и условного порядкового номера. Идентичность рассказчиков обозначена при помощи специальных кодов. Код состоит из инициалов, указания на пол и возраст.

028-Kibrick_AA:_Layout 1 18.05.2008 15:13 Page 188

тональный диапазон	типичный тон начал (ресетов)	целевой уровень финальных падений	целевой уровень нефинальных паде-ний	целевой уровень подъемов при запя-тых	уровень заударного падения при запятых	уровень подъема при многоточиях	другие характери-стики
190 – 500+	240-290 (первый 310)	190-200	220, часто с загибом	400-500, однажды 300	220-240	280-360	Хорошо выраженные дви-жения тона.. Упередненная артикуляция. Говорит немного в нос.

Таблица 2. Просодический портрет рассказчика рассказа Z35 (АМ м 9 лет)

тональный диапазон	типичный тон начал (ресетов)	целевой уровень финальных падений	целевой уровень нефинальных паде-ний	целевой уровень подъемов при запя-тых	уровень заударного падения при запятых	уровень подъема при многоточиях	другие характери-стики
210-480	240-280	200	220-240	370-470	220-240	260-340	Хорошо выраженные движения тона.. Упередненная артикуляция. Говорит немного в нос. Часто скрипучий голос на гласных. Почти все идет на многоточиях

Таблица 3. Просодический портрет рассказчика рассказа Z36 (АМ м 9 лет)

тональный диапазон	типичный тон начал (ресетов)	целевой уровень финальных падений	целевой уровень нефинальных паде-ний	целевой уровень подъемов при запя-тых	уровень заударного падения при запятых	уровень подъема при многоточиях	другие характери-стики
200 – 500+	280-320	200?	210-250	360-480, однажды 330	230-240	—	В этом рассказе странно высокий уровень ресетов по сравнению с другими у этого рассказчика

Таблица 4. Просодический портрет рассказчика рассказа Z38 (АМ м 9 лет)

тональный диапазон	типичный тон начал (ресетов)	целевой уровень финальных падений	целевой уровень нефинальных паде-ний	целевой уровень подъемов при запя-тых	уровень заударного падения при запятых	уровень подъема при многоточиях	другие характери-стики
220-400	260-290	надежных дан-ных нет	240-250	330-400	260-270	290-330 (строки 6-7 – 400Гц?)	Хриплый голос

Таблица 5. Просодический портрет рассказчика рассказа Z39 (АМ м 9 лет)

028-Kibrick_AA:_Layout 1 18.05.2008 15:13 Page 189

тональный диапазон	типичный тон начал (ресетов)	целевой уровень финальных падений	целевой уровень нефинальных паде-ний	целевой уровень подъемов при запя-тых	уровень заударного падения при запятых	уровень подъема при многоточиях	другие характери-стики
220-400	250-290	данных нет	220	330-400	240-300	–	Часто в паузах шмыгает носом. К концу уро-вень подъемов в запятых заметно снижается: вначале около 400 Гц, в конце 340, и к тому же тихо.

Таблица 6. Просодический портрет рассказчика рассказа Z40 (АМ м 9 лет)

тональный диапазон	типичный тон начал (ресетов)	целевой уровень финальных падений	целевой уровень нефинальных паде-ний	целевой уровень подъемов при запя-тых	уровень заударного падения при запятых	уровень подъема при многоточиях	другие характери-стики
190 – 500+	260-290	данных нет	220	370-410	220-260	340-380, однажды 420	

Таблица 7. Просодический портрет рассказчика рассказа Z41 (АМ м 9 лет)

Учитывая все эти данные, можно построить объединенный просодический портрет рассказчика, гораздо более надежный, чем в случае лишь единичного рассказа.

тональный диапазон	типичный тон начал (ресетов)	целевой уровень финальных падений	целевой уровень нефинальных паде-ний	целевой уровень подъемов при запя-тых	уровень заударного падения при запятых	уровень подъема при многоточиях	другие характери-стики
190 – 500+	240-290, редко до 320	190-200	210-250, иногда с загибом	330-500, редко 300	220-270	260-360, редко до 420	Упередненная артикуляция. Говорит немного в нос, часто хрипло. Часто скрипучий голос на гласных. Часто использует интонацию многоточия

Таблица 8. Суммарный просодический портрет рассказчика АМ (м 9 лет)

Рассказчик АМ имеет большой тональный диапазон. Верхняя часть этого диапазона реализуется в случаях ухода в фальцет. Базовый уровень начал ЭДЕ составляет 240-290 Гц, очень редко выходя за пределы этой полосы. Полоса финального падения узкая и четко фиксированная: 190-200 Гц. Нефинальные падения хорошо отличимы от финальных: 210-250 Гц. В тех случаях, когда различие небольшое, говорящий часто пользуется загибом частотной кривой вверх на заударных слогах. При интонации запятой рассказчик уходит в верхнюю часть голосового диапазона, иногда в фальцет. Целевой уровень заударного падения при запятых очень близок к уровню нефинальных падений. Целевой уровень при многоточиях пересекается с диапазоном подъема при интонации канонической запятой. В таких случаях интонация многоточия отличается от интонации запятой благодаря ряду других признаков – удлинению акцентированной гласной, сложному восходяще-ровному тону, медленному характеру подъема, отсутствию заударного падения в низкий уровень и нек.др.

028-Kibrick_AA:_Layout 1 18.05.2008 15:13 Page 190

6. Применение просодических портретов

Если бы реализация просодических звукотипов была элементарной, то роль просодических портретов сводилась бы просто к перечню характеристик того или иного говорящего. Это позволило бы решить относительно простую проблему: люди имеют разный тональный диапазон и проводят семиотические различия внутри него по-разному. Например, то, что для одного является типичным целевым уровнем подъемов при запятой, для другого является уровнем, характерным для интонации многоточия.

Более сложной является другая проблема. В речи каждого говорящего есть более и менее прототипические инстанции одного и того же просодического звукотипа. Как уже отмечалось выше, семиотически противопоставленные звукотипы в отдельных случаях могут нейтрализоваться. Просодические портреты помогают выявить базовую, исходную семиотическую систему данного говорящего и анализировать более сложные случаи уже на ее основе.

Кроме того, многие просодические паттерны конституируются не одной акустической характеристикой, а целым их набором – ср. список компонентов интонации многоточия в конце раздела 5. В одних случаях оказываются более значимыми одни из этих компонентов, в других – другие. Просодические портреты помогают обнаружить значимые компоненты в каждом отдельном случае и опереться при определении паттерна на твердую эмпирическую основу.

Пример дискурсивной транскрипции (рассказ Z46, рассказчица КЖ ж 14 лет)

11. ...(0.8) /Потом ..(0.2) нам встретился <тут | какой-то> /мостик,
12. ...(0.6) очень ..(0.1) ’ /\узенький,
13. ...(0.7) <и> /мы через него еле /проехали тоже,
14. ....(1.2) /потом подъехали к \/пляжу,
15. ...(0.5) (/Я вообще плавать не \уме-ею.
16. ..(0.4) Не \умела тогда,
17. *когда мне это \снилось.*)
18. ....(1.1) \вот,
19. и-и’ ....(1.1) /я почему-то \поплыла.
20. /Нырнула,
21. и \поплыла.

Строки обозначают номера ЭДЕ в рамках рассказа. Случаи интонации точки представлены в строках 15, 17, 19 и 21. Интонация запятой с падением – в строках 12, 14, 16, 18. Обоснованное различение двух типов падений стало возможно в этом рассказе – как и во многих других случаях – лишь при помощи построения просодического портрета данной рассказчицы и выяснения семиотических противопоставлений, используемых ею в частотном континууме. В строках 11, 13 и 20, наконец, можно видеть каноническую интонацию запятой с подъемом тона в главном акценте.

7. Заключительные замечания

Несомненно, представленный здесь формат просодического портрета говорящего – лишь самый первый эскиз методологического инструмента. Более детальная разработка этого инструмента – дело дальнейших исследований. К примеру, раздел «прочие характеристики» должен быть расщеплен на ряд значимых позиций, характеризующих разные слои просодии.

Возможный контраргумент против предложенного здесь подхода состоит в том, что транскрибирование устного дискурса оказывается слишком трудоемким процессом. При этом подходе исследователь вынужден осуществлять «нулевой цикл» анализа, который предшествует собственно транскрибированию и на котором изучается система данного говорящего. Возразить на это нечего, кроме того, что такова реальность. Транскрибирование устного дискурса – это действительно трудо- и времяемкий процесс. Затраты усилий отчасти окупаются тем, что получается продукт более высокой ценности.

028-Kibrick_AA:_Layout 1 18.05.2008 15:13 Page 191

Список литературы

1. Кибрик А.А., Подлесская В.И. 2003. К созданию корпусов устной русской речи: принципы транскрибирования // Научно-техническая информация. Серия 2, 6. С. 5-11.

2. Кибрик А.А., Подлесская В.И. 2006. Проблема сегментации устного дискурса и когнитивная система говорящего // Соловьев В.Д. (ред.) Когнитивные исследования. Вып. 1. М.: Институт психологии РАН. С. 138-158.

3. Кибрик А.А., Подлесская В.И. (ред. 2009.) Семантика и структура устного дискурса (на материале корпуса рассказов о сновидениях). М: ЯСЛ (в печати).

4. Кодзасов С.В. 2002. Фазовая символика тона // Арутюнова Н.Д. (ред.) Логический анализ языка. Семантика начала и конца. М.: Индрик.

5. Кривнова О.Ф. 2007. Ритмизация и интонационное членение текста (опыт теоретико-экспериментального исследования). Диссертация на соискание ученой степени доктора филологических наук. М.: МГУ им. М.В. Ломоносова.

6. Светозарова Н.Д., Вольская Н.Б., Павлова А.В., Шитова Л.Ф. 1988. Просодическая организация русской спонтанной речи // Светозарова Н.Д. (ред.) Фонетика спонтанной речи. Л.: Изд-во Ленинградского университета. - 141-182.

7. Хитина М.В. 2004. Делимитативные признаки устно-речевого дискурса. М.: МГЛУ.

8. Chafe Wallace. 1994. Discourse, consciousness, and time. Chicago: University of Chicago Press.

029-Kobzareva_T:_Layout 1 18.05.2008 14:52 Page 192

ПОСТРОЕНИЕ ГРАФА СВЯЗЕЙ СЕГМЕНТОВ[1] (ПОВЕРХНОСТНО-СИНТАКСИЧЕСКИЙ АНАЛИЗ РУССКОГО ПРЕДЛОЖЕНИЯ)

BUILDING A GRAPH OF SEGMENT CONNECTIONS (RUSSIAN SENTENCE SURFACE-SYNTACTICAL ANALYSIS)

Кобзарева Т.Ю. *(stamstam@mtu-net.ru)*
Российский государственный гуманитарный университет

В работе обсуждается лингвистический базис построения связей сегментов в русском предложении и некоторые проблемы, возникающие при поиске слова – хозяина сегмента.

1. Введение

Рассматривается лингвистический базис модуля построения графа связей сегментов в русском предложении в системе поверхностно-синтаксического анализа (ПСА), разрабатываемой в настоящее время в РГГУ.

В системе вся процедура анализа делится на модули [4]. Каждый из модулей на основе минимально необходимой словарной информации, приписываемой словоформе при морфанализе, грамматической информации, полученной на предшествующих этапах, и идеальной модели рассматриваемых явлений обрабатывает определенную зону синтаксических явлений. Система выстроена как лингвистически оптимальная последовательность процедур анализа и способов их организации, которые позволяют минимизировать используемую на каждом этапе информацию и упростить каждый очередной этап работы с текстом.

2. Используемые понятия

Используется следующая идеальная модель русского предложения (S) [1,2]. S состоит из цепочки **β-сегментов –** простых-главных предложений; каждый из **β-сегментов** может быть разорван вложением в него **α-сегментов -** придаточных предложений и обособленных оборотов (деепричастных оборотов, согласованных определений, предложных, вводных и сравнительных оборотов). Каждый **α-сегмент** в свою очередь может быть разорван следующего уровня вложениями в него **α-сегментов,** причем количество вложений одного уровня, как и количество уровней вложений теоретически ничем не ограничено.

Границами сегментов являются знаки препинания, эксплицитно задаваемые правилами русской пунктуации, и, реже, – сочинительные союзы.

3. Иерархия процедур анализа

Вся процедура анализа делится на следующие модули [4]:

1. **Постморфология** – решение несловарных проблем морфанализа [5];
2. **Разрешение омонимии частей речи** [6].
3. **Предсегментация** – построение проективных фрагментов определительных именных и предложных групп (ИГ и ПГ) [7], сложного сказуемого и т.д., выступающих единицами анализа в модуле Сегментации [2,4];
4. **Сегментация** – построение сегментов (придаточных предложений, деепричастных и других обособляемых оборотов и простых–главных предложений в составе сложных и сложноподчиненных предложений): определение их левых и правых границ и одновременное элиминирование разрывов, возникающих при вложении сегментов в сегменты [2,3,4].
5. **Внутрисегментный анализ** – поиск связей слов внутри построенных сегментов [4].
6. **Межсегментный анализ** – построение связей между уже построенными сегментами – графа связей сегментов.

К этапу построения графа связей сегментов построены сегменты и построены все сочинительные и подчинительные связи слов внутри сегментов, определяющие проективные фрагменты [1,2,3,7], в частности -

[1] Доклад подготовлен при частичной поддержке РФФИ (грант 06-06-80434).

029-Kobzareva_T:_Layout 1 18.05.2008 14:52 Page 193

зоны влияния сочиненных предикатов [8], что используется в модуле межсегментного анализа.

При этом для каждого сегмента, потенциально имеющего хозяина, решается два вопроса: какое слово может быть его хозяином и где это слово-хозяин искать.

Для русского предложения, потенциально имеющего весьма сложную сегментную структуру и большую длину, целесообразно, а главное – возможно, как показывают теоретические и экспериментальные данные, строить сегменты до полного анализа внутрисегментных и межсегментных связей [2,4]. После модуля сегментации, который при построении сегментов снимает омонимию знаков препинания и сочинительных союзов (т.е. все сочинительные связи строятся в модуле сегментации при определении функций знаков препинания) [2,3], работает модуль внутрисегментного анализа, строящий все подчинительные связи внутри сегментов [11], и только после этого – модуль межсегментного анализа.

4. Отношения между сегментами

При межсегментном анализе мы исходим из того, что между сегментами могут быть два вида отношений: отношения подчинения и сочинения.

1. Отношение подчинения выражается в том, что некоторое слово сегмента-хозяина является хозяином некоторого слова сегмента-слуги, это значит, что каждый конкретный случай отношения подчинения сегментов манифестируется подчинительной связью двух слов из разных сегментов.

2. Под отношением сочинения понимаем два вида отношений сегментов.

2.1. Отношение сочинения при соподчинении, когда два α-сегмента подчинены одному слову сегмента-хозяина. Это отношение выражается в сочинении двух слов разных α-сегментов одного класса – придаточных, деепричастных оборотов или определительных оборотов. Для обособленных оборотов это – сочинение их вершин, а для придаточных – подчинительных союзов (п\с).

2.2. Отношение простых-главных предложений – β-сегментов, входящих в одно предложение, традиционно рассматриваемое как их сочинение и выражаемое как сочинение их предикатных вершин.

Все связи сегментов предложения представляются при этом в виде графа связей сегментов, где каждая связь манифестируется связью двух слов разных сегментов.

При этом каждый α-сегмент потенциально имеет хотя бы одного хозяина и для каждого α-сегмента при поиске его хозяина решается два вопроса:

1. что́ именно искать, т.е. какое слово в принципе может быть его хозяином и

2. где искать, т.е. в какой зоне предложения может находиться слово – его хозяин.

При этом определение слова-слуги в межсегментных связях не составляет проблемы: у α-сегментов – оборотов в роли слуги слово-слуга всегда выступает вершина сегмента, а в придаточных слуга – подчинительный союз.

5. Особенности сегментной структуры предложения, осложняющие поиск межсегментных связей

Рассмотрим некоторые факторы, которые осложняют поиск хозяина, т.е. ответа на вопросы что́ и где́ искать.

5.1. Факторы, осложняющие ответ на вопрос, что́ искать, т.е. какое слово является хозяином α-сегмента.

Придаточные предложения (ПП) способны:

1) заполнять валентности предикативных слов и

2) выступать в роли слуг актантов и сирконстантов – существительных в качестве присубстантивных определений или уточнений предложных групп и наречий.

Соответственно, при поверхностно-синтаксическим анализе для каждого ПП надо 1) представлять себе, что потенциально может быть хозяином такого ПП и 2) найти его хозяина в данном контексте.

При этом в зоне поиска может оказаться несколько слов, претендующих на роль слова-хозяина. Соответственно, нужно предусматривать, что может возникать синтаксическая неоднозначность межсегментных связей.

5.2. Факторы, осложняющие ответ на вопрос, где́ искать хозяина α-сегмента.

5.2.1. Подчиненные сегменты, «вложенные» по смыслу в сегмент-хозяин, могут физически не иметь в линейной структуре контакта с сегментом-хозяином.

S1 *Иван, зная все это,* ***обзавелся*** *двумя вязками бубликов и колбасою* ***и****, спросивши рюмку водки, в которой не бывало недостатка ни в одном постоялом дворе,* ***кончил*** *свой ужин, усевшись на лавке перед дубовым столом, вкопанным в глиняный пол* (Набоков)

029-Kobzareva_T:_Layout 1 18.05.2008 14:52 Page 194

Деепричастный оборот *спросивши рюмку водки* отделен от своего хозяина-сказуемого ***кончил*** придаточным *в которой не бывало недостатка ни в одном постоялом дворе.*

S2 *Я* ***узнал*** *также,* ***что*** *его здоровье слабо,* ***что*** *жизненный пыл, угасший в этом человеке, оставил его тело без присмотра и без поощрения,* ***что*** *он скоро умрет*» (Набоков)

ПП = *что жизненный пыл оставил его тело без присмотра и без поощрения* подчинено сказуемому простого-главного ***узнал***, но отделено от сегмента-хозяина ПП = *что его здоровье слабо,* а между последним придаточным *что он скоро умрет* и сегментом-хозяином находится ПП = *что его здоровье слабо* и матрешка – ПП с вложенным в него обособленным А-оборотом *что жизненный пыл, угасший в этом человеке, оставил его тело без присмотра и без поощрения*.

5.2.2. α-сегменты могут 1) быть соподчинены одному хозяину или 2) образовывать цепочку α-сегментов, связанных последовательно.

S3. *В тот* ***час****, когда и сил не было дышать, когда солнце, раскалив Москву, в сухом тумане валилось куда-то за Садовое кольцо, - никто не пришел под липы, никто не сел на скамейку, пуста была аллея.* Два присубстантивных ПП служат определениями слова ***час***, т.е. соподчинены и сочинены.

S4. *И вот в то* ***время****,* ***когда*** *он рассказывал поэту о* ***том****,* ***как*** *ацтеки лепили из теста фигурку, в аллее показался первый человек* придаточные образуют цепочку последовательно подчиненных: (*в то*) ***время*** R ***когда*** и (*о*) ***том*** R ***как***.

К этапу построения графа связей сегментов построены сегменты и построены все сочинительные и подчинительные связи слов внутри сегментов, определяющие **все** проективные фрагменты линейной структуры, в частности, - зоны влияния сочиненных предикатов, что и используется в модуле межсегментного анализа.

6. Проективность межсегментных связей

Узлами графа межсегментных связей являются сегменты. Их связи выражаются связью двух слов разных сегментов. Из свойства проективности связей, введенного Теньером и обычно рассматриваемого на примерах связей слов внутри одного сегмента [9], естественным образом вытекает свойство проективности межсегментных связей. В редких специфических случаях проективность межсегментных связей может нарушаться (см. Рис.1).

S5. *Бородатые студенты в клетчатых пледах, смешавшись с жандармами в пелеринах, предводительствуемые козлом регентом, в буйном восторге выводя, как плясовую, вечную память, вынесут полицейский гроб с останками моего дела из продымленной залы окружного суда.* (Мандельштам)

Рис.1

Свойство проективности межсегментных связей позволяет ограничить зону поиска хозяина а-сегмента-слуги. Для этого в модуле межсегментного анализа при поиске хозяина a^k-сегмента используются три общих правила проективности межсегментных связей.

6.1. Общие правила проективности межсегментных связей.

Правило 1. Если α^k-сегмент полностью погружен в α^j-сегмент, т.е. в предложении **и левее, и правее** α^k-сегмента есть хотя бы одно слово (включая границы – сочинительные союзы или сочинительные союзы в составе многокомпонентных границ, состоящих из знаков препинания и слов) α^j-сегмента (м.б. неконтактно!), то хозяин α^k-сегмента и сочиненный-соподчиненный ему α-сегмент не могут находиться в линейной структуре предложения ни левее, ни правее границ α^j-сегмента.

Правило 2. Если **левее (или правее)** α^k-сегмента есть часть (хотя бы одно слово) **β-сегмента** (м.б. неконтактно!), то хозяин α^k-сегмента и, соответственно, сочиненный-соподчиненный ему α^i-сегмент не могут находиться левее (или, соответственно, правее) этой – ближайшей – слева (или справа) к α^k-сегменту – части β-сегмента.

Правило 3. Если **правее** некоторого отрезка a^k-сегмента есть левая граница a^i-сегмента (м.б. неконтактно!), то хозяин a^k-сегмента (и сочиненного и соподчиненного ему a^n-сегмента) не может находиться в линейном отрезке предложения между левой и правой границами a^i-сегмента.

029-Kobzareva_T:_Layout 1 18.05.2008 14:52 Page 195

S6 *Оглядываясь на девятнадцатый век русской культуры,* **[*разбившийся, конченый, неповторимый*], [*которого никто не смеет повторять*],** *я* ***вижу*** *в нем единство непомерной стужи, спаявшей десятилетия в один денек, в одну ночку, в глубокую зиму, где страшная государственность - как печь, пышущая льдом*

6.2. Поиск хозяина деепричастного оборота *(D^v-об).*

Хозяином деепричастия (D^v) является предикат сегмента-хозяина. Поэтому для D^v-об, кроме перечисленных ограничений, важно определить, в зоне влияния **какого именно предиката** он находится. При построении сегментов строятся все сочинительные связи внутри сегментов, и важной составляющей анализа при построении придаточных и простых-главных предложений является анализ сочинения предикатов, в ходе которого определяются знаки препинания и союзы, являющиеся границами зон влияния предикатов, т.е. частей линейной структуры предложения, где могут находиться слуги каждого из предикатов.

Соответственно, при поиске хозяина D^v-об из трех правил проективности межсегментных связей выводятся следующие условия:

Условие 1. При полном погружении D^v-об в сегмент его хозяином является предикативная вершина (ПВ) сегмента, в зоне влияния которой находится D^v-об: при этом границы зон влиянии ПВ – это границы сегмента, в который погружен D^v-об или, при сочинении нескольких ПВ, сужающие границы – операторы, сочиняющие ПВ.

S7 *Она поднималась по седым ступеням; от нее* ***шло*** *знакомое тепло,* $\boldsymbol{и^n}$*,* ***поднимаясь*** *мыслью рядом с ней, я* ***видел*** *нашу предпоследнюю встречу на званом вечере в парижском доме, где было очень много народу,* $\boldsymbol{и^k}$ *мой милый друг,* ***желая*** *мне оказать какую-то тонкую эстетическую услугу,* ***тронул*** *меня за рукав* $\boldsymbol{и^{k+1}}$ *подвел меня к Нине,* ***сидевшей*** *в углу дивана,* ***сложившись*** *зетом, с пепельницей у каблучка* (Набоков).

D^v-об = ***поднимаясь*** *мыслью рядом с ней* оказывается в зоне влияния предиката ***видел***, ограниченной слева сочинительным союзом $\boldsymbol{и^n}$**,** и строится связь ***видел*** **R** ***поднимаясь;*** D^v-об = ***желая*** *мне оказать какую-то тонкую эстетическую услугу* полностью погружен в простое-главное = $\boldsymbol{и^k}$ *мой милый друг* ***тронул*** *меня за рукав* $\boldsymbol{и^{k+1}}$ *подвел меня к Нине* с $\boldsymbol{и^k}$ - левой границей зоны влияния ***тронул*** и правой ее границей $\boldsymbol{и^{k+1}}$, из чего получаем связь ***тронул*** **R** ***желая.*** D^v-об = ***сложившись*** *зетом* полностью погружен в сегмент ***сидевшей*** *в углу дивана с пепельницей у каблучка* с единственной ПВ= ***сидевшей***, следовательно строится ***сидевшей*** **R** ***сложившись.***

S8. *Затем* ***выходил*** *александрийский актер* ***и***, ***ударяя*** *себя в грудь, истошным голосом,* ***закатываясь*** *от крика и* ***переходя*** *в зловещий шепот,* ***читал*** *стихотворение Никитина "Хозяин".* (Набоков)

В S8 все три D^v-об полностью погружены в β-сегмент *Затем* ***выходил*** *александрийский актер* ***и*** ***читал*** *стихотворение Никитина "Хозяин".* Границей зон влияния предикатов является сочиняющий их союз ***и***. Все три D^v-об находятся в зоне влияния ***читал***, и, соответственно, ему подчинены.

Условие 2 (вытекает из 2-ого и 3-ого общих правил проективности и Условия1 поиска хозяина D^v-об): если D^v-об – первый сегмент в предложении или непосредственно слева – двоеточие или точка с запятой, то его хозяин – первый предикат ближайшего справа от него β-сегмента.

S9. ***Поднявшись*** *по лестнице, мы* ***очутились*** *на щербатой площадке; отсюда видна была гора с собранием крапинок костяной белизны на боку;* ***огибая*** *подножье,* ***бежал*** *дымок невидимого поезда и вдруг скрылся.*(Н)

В **S9** есть два D^v-об, непосредственно справа к каждому примыкает β-сегмент. По Условию 2 строятся соответственно связи ***очутились*** R ***Поднявшись*** и ***бежал*** R ***огибая***.

Рассмотрим Условие2 в случае, осложненном вставлением α-сегмента или матрешки из α-сегментов между D^v-об и ближайшим справа фрагментом β-сегмента:

S10 ***Оглядываясь*** *на девятнадцатый век русской культуры,* **[*разбившийся*], [*конченый*], [*неповторимый*], [*которого никто не смеет повторять*],** *я* ***вижу*** *в нем единство непомерной стужи, спаявшей десятилетия в один денек, в одну ночку, в глубокую зиму, где страшная государственность - как печь, пышущая льдом.* (Набоков)

Для D^v-об = ***Оглядываясь*** *на девятнадцатый век русской культуры* Условие 2 позволяет легко найти ***вижу*** – хозяина D^v-об: 1) хозяин анализируемого D^v-об (*Оглядываясь на…)* по 3-ому общему условию проективности не может быть в α-сегментах справа от D^v-об; 2) хозяин D^v-об по 2-ому общему правилу проективности не может быть правее ближайшего слева или справа β-сегмента (здесь есть только β-сегмент справа), т.е. хозяин м.б. только в β-сегменте *я* ***вижу*** *в нем единство непомерной стужи.* Так как в нем есть ровно один предикат ***вижу***, строится связь ***вижу*** R ***Оглядываясь***.

7. Специфика поиска хозяина ПП

Общие правила проективности определяют зону поиска хозяина ПП, т.е. сегмент или зону влияния предиката, где может находится хозяин ПП. Но для ПП встает вопрос, какое именно слово сегмента-хозяина может быть хозяином ПП в силу уже упомянутой потенциальной функциональной неоднозначности ПП.

029-Kobzareva_T:_Layout 1 18.05.2008 14:52 Page 196

Рассмотрим синтаксически возможные пути решения этой проблемы.

7.1. Виды ПП.

ПП по подчинительному союзу могут относиться к одному из трех видов.

А. ПП, выступающие в подавляющем числе случаев в роли присубстантивных, т.е. определительных ПП с подчинительным союзом ***который***.

Б. ПП, способные выступать только в роли слуг предикативных вершин сегмента-хозяина. Это, например, ПП с подчинительными союзами ***если*** и ***хотя.***

В. ПП, способные и заполнять валентности предикатов, и быть определениями слуг предикатов, в частности присубстантивными. К таким ПП относятся, например, ПП с подчинительными союзами ***кто, что, где, когда, как...*** :

…решает*, **что** книга бездарна,… vs. ..**город**, **что** на горе…*

…знаю, как*…vs. …показателем **того, как** развивалось*

…знаю*, **где** он спрятал… vs. …в десятой **главе, где** автор…*

7.1.1. А.

Кратко рассмотрим определительные ПП с ***который***. При поиске хозяина ***который*** мы можем использовать особенности его согласования со словом-хозяином (правила конгруэнтности), а именно: согласование по числу с учетом потенциального сочинения слов-хозяев, согласование в единственном числе по роду и, в некоторых случаях дающее возможность избежать синтаксической неоднозначности, согласование по одушевленности (для ***который*** в Вин.п.м.р.ед.ч. и в Вин.п.мн.ч.): ***дом*** *отца, на* ***который*** *не могу смотреть * ***книги*** *сестер, на* ***которые*** vs. *дом* ***отца****, на* ***которого*** *не могу смотреть\ книги* ***сестер****, на* ***которых****.*

7.1.2. В.

Рассмотрим возможные пути разрешения функциональной неоднозначности ПП при поиске хозяина для ПП группы В. При этом мы действуем методом исключения: сначала проверяем, нет ли у ПП хозяина, не являющегося ПВ.

7.1.2.1. Скрепы.

Для связи сегмента-хозяина и ПП – его слуги часто используются слова, которые называют скрепами. Скрепы в своем сегменте уже имеют хозяина, и они объявляется хозяином подчинительного союза. Присубстантивными при наличии скреп-хозяев считаем ПП, скрепы которых выступают в функции существительных.

S11. *Хлестаков порхает по пьесе, не желая толком понять, какой он поднял переполох, и жадно стараясь урвать **все**, **что** подкидывает ему счастливый случай.*(Н) Скрепа – ***все*** **–**местоименное существительное.

S12. *Беда в **том**, **что** ни искренность, ни честность, ни даже доброта сердечная не мешают демону пошлости завладеть пишущей машинкой автора.* (Н) Скрепа ***(в) том*** **–** тоже местоименное существительное.

Связи ***все*** R ***что*** и ***(в) том*** R ***что*** определяют соответствующие придаточные как присубстантивные.

Часто скрепа является согласованным определением существительного, и тогда она маркирует существительное, являющееся хозяином подчинительного союза (п\с). В этой ситуации ПП тоже выступают в определительной функции, а хозяином п\с является существительное – хозяин скрепы.

S13. *Он находился **в том <u>состоянии</u>** чувств и души, **когда** существенность, уступая мечтаниям, сливается с ним в неясных видениях первосонья.*(Н) Местоименное прилагательное - скрепа ***тот*** маркирует хозяина присубстантивного ПП: (***в том***) ***<u>состоянии</u>*** R ***когда.***

Скрепы	**п\с в α^k**
Сп.1: там, там-то, туда, туда-то, оттуда, оттуда-то, везде, всюду, повсюду, нигде, никуда, где-то, кое-где… Сп.2 (скрепы присубстантивные): тот, каждый, какой-нибудь, какой-то, следующий ...	где, куда, откуда
так, такой, таков, столько, настолько, постольку, тот	как
такой, таков	какой
столько, настолько, так, столь,	сколько, насколько
Сп.1: так, настолько, столь, столько, что-то, потому, оттого, таков, Сп.2 (присубстантивные): тот (мн.ч. или ед.ч.ср.р.), такой, нечто, ничто, кое-что, что-нибудь, такой, один, все (мн.ч. или ср.р. ед.ч.), первый, единственный, другой, весь, последний…	что
постольку, такой, таков	поскольку
тот, такой, всякий, каждый, некоторый, многие, любой, первый, последний, единственный, все(мн.ч.), никто, кто-то, кто-нибудь, кое-кто, некто, один…	кто, чей
иначе, по-другому, по-иному, иной, другой…	чем, нежели
так, затем, настолько, (не) такой, таков, столько, настолько, тот, слишком, столь…	чтобы

029-Kobzareva_T:_Layout 1 18.05.2008 14:52 Page 197

Скрепы	**п\с в α^k**
тогда, всегда, иногда, изредка, когда-нибудь, когда-то, кое-когда, тот (слуга)	когда
потому	почему
затем	зачем

Таблица 1. Примеры открытых списков скреп для некоторых п\с

Если скрепы при поиске хозяина присубстанивного ПП нет, при определении хозяина для многих ПП, например, с союзами ***где*** и ***когда,*** возникают трудности.

При этом на втором шаге (когда скрепы нет) для ***где*** и ***когда*** можно использовать предлоги.

Если в части сегмента-хозяина слева от сегмента-слуги (с определенными ограничениями на то, какие слова могут быть между ними) есть предложная группа с предлогом, который может иметь значение места – ***в\на\к\под\за\над…*** и существительное – его слуга - предмет неодушевленный, велика вероятность, что это существительное и есть хозяин ***где.*** Аналогично, если в части сегмента-хозяина слева от сегмента слуги есть ПГ с предлогом ***до \ перед \ после \ в течение \в продолжении *** ... , существительное – слугу в этой ПГ мы можем считать хозяином ***когда.***

S14. *<u>В тот самый день</u>,* ***когда*** *происходила всякая нелепая кутерьма, вызванная появлением черного мага в Москве, <u>в пятницу,</u>* ***когда*** *был изгнан обратно в Киев дядя Берлиоза,* ***когда*** *арестовали бухгалтера*...(Булгаков)

В ситуациях, когда при поиске хозяина функционально неоднозначных п\с не найдено ни скреп, ни предложных групп с соответствующим значением, и в сегменте-хозяине нет ситуации, когда ПП – единственный претендент на заполнение валентности предиката, необходимо понять, какие слова могут претендовать на роль хозяина соответствующего п\с. При этом не удается обойтись без использования «семантических» классов.

Хозяином ***когда*** могут быть 1) существительные, называющие или подразумевающие некоторый момент или интервал времени или называющие некоторые явления, имеющие ограниченную протяженность во времени: *январь, февраль, март, апрель, май, июнь, июль, август, сентябрь, октябрь, ноябрь, декабрь; зима, весна, лето, осень; понедельник, вторник, среда, четверг, пятница, суббота, воскресенье время, век, год, столетие, день, час, минута, период..;. урок, лекция, эксперимент, выступление, поездка, доклад, сон ...; момент, финал,* ***…;*** или 2) наречия времени: *весной, вечером, впоследствии, давно, днем, изредка, зимой, летом, накануне, недавно, некогда, ночью, осенью, позже, потом, раньше, сначала, утром…*

Сложнее определить семантические классы слов – хозяев ***где.***

Если нет соответствующей скрепы, хозяином – не предикатом – могут быть наречия места: *внизу, наверху, посередине, слева, справа, наверху, внизу, сбоку…*

или некоторые объекты, условно – «вместилища», для которых возможна конструкция «в\на Х, где». К таким относятся строения, их части, предметы, имеющие объем – физические вместилища: *дом, комната, дача, имение, особняк, пансион, помещение, город, музей, кабинет, проход, пустырь, музей, заведение, веранда, теплица, здание ...;пещера, туннель, подземный переход, …*; *коробка, шкаф, полка…;* физические объекты, в одном из своих значений мыслимые как имеющие объем и\или поверхность: *небо, река, море, гора, стол, дорога, озеро…*; географические объекты типа *район, страна, город, место, мир,...* ; объекты, имеющие поверхность: *карта, лист, крыша, стена, картина…*; объекты, могущие выступать как физические и\или как «ментальные вместилища»: *книга* (и все объекты типа книги: *журнал, задачник, сборник, энциклопедия, словарь, учебник…) статья, доклад, рассказ, сочинение, обзор, план, текст, абзац, глава…*

Трудным для поиска хозяина является и п\с ***что***:

чаще всего в присубстантивных ПП он заменяет ***который*** или является слугою существительного, семантически связанного с предикатом некоторого ментального действия: *мысль, понимание, слова, предположение, подозрение, предчувствие, подтверждение, рассуждение, слух, страх, убеждение, вероятность, ответ, возможность, донесение, допущение, знание, известие, открытие, условие ...*

Таким образом, мы видим, что для построении графа сегментных связей мы нуждаемся в словарной информации о способности существительных присоединять придаточные с определенными п\с. Способность эта определяется семантикой существительных и нуждается в отдельном исследовании.

Если мы можем утверждать, что существительных – претендентов на роль хозяина определенного п\с в сегменте-хозяине нет, то ПП, вероятнее всего, является слугою той ПВ сегмента, в зоне влияния которой ПП находится.

029-Kobzareva_T:_Layout 1 18.05.2008 14:52 Page 198

8. Заключение

Рассмотрены некоторые общие и частные проблемы построения графа поверхностно-синтаксических связей сегментов и подход к их решению, который может быть использован при анализе русских предложений любой сегментной структуры. Алгоритмы модуля межсегментного анализа в настоящее время программируются на разработанном в РГГУ языке объектного моделирования [10].

Опыт построения лингвистического базиса модуля межсегментного анализа показывает, что для анализа этого уровня необходима семантическая информация о способности слов присоединять ПП с определенными п\с.

Вышеизложенным не исчерпывается, естественно, весь круг проблем, возникающих при межсегментном анализе, однако алгоритмы, построенные на этой базе, уже позволяют решать достаточно сложные задачи этого уровня анализа.

Список литературы

1. Кобзарева Т.Ю. Некоторые аспекты анализа сочинения при сегментации русского предложения // КИИ'2002. Труды восьмой национальной конференции по искусственному интеллекту с международным участием. М. Физматлит. Т.1.С. 192-198.

2. Кобзарева Т.Ю. Принципы сегментационного анализа русского предложения // Московский лингвистический журнал. М. 2004. Т.8 №1, с. 31-80.

3. Кобзарева Т.Ю. Омонимия и синонимия знаков препинания в русском тексте // Труды Международной конференции Диалог'2005. — М.: Наука,2005 — С. 233-237.

4. Кобзарева Т.Ю. Иерархия задач поверхностно-синтаксического анализа русского предложения // НТИ, Сер.2, №1, 2007, с 23 – 35.

5. Кобзарева Т.Ю. Морфанализ in vivo. // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2004. — М.: Наука, 2004. — С 286-291.

6. Кобзарева Т.Ю., Афанасьев Р.Н. Универсальный модуль предсинтаксического анализа омонимии частей речи в русском языке на основе словаря диагностических ситуаций // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции Диалог'2002 в 2 т. Т.2. — М.: Наука, 2002. — С 258-268.

7. Кобзарева Т.Ю. Некоторые свойства линейной структуры именных и предложных групп (Поверхностно-синтаксический анализ русского предложения)// Вестник РГГУ. № 8/07, Серия «Языкознание» (Московский лингвистический журнал № 9/2), Москва 2007. – С. 113-130.

8. Кобзарева Т.Ю. Рекурсивность и проективность сочинительных связей в русском тексте // Компьютерная лингвистика и интеллектуальные технологии Труды Международной конференции Диалог 2006, Бекасово, 31 мая – 4 июня 2006 г. — М.: Наука, 2006. — С. 223-229.

9. Теньер Люсьен, Основы структурного синтаксиса. — М.: Прогресс, 1988.

10. Баталина А.М., Епифанов М.Е., Ивличева О.О., Кобзарева Т.Ю., Лахути Д.Г. Инструментальная среда для экспериментов с алгоритмами поверхностно-синтаксического анализа // Труды Международной конференции Диалог'2004, — М.: Наука, 2004 — С. 32-38.

11. Баталина А.М. Епифанов М.Е. Кобзарева Т.Ю. Кушнарёва Е.В. Лахути Д.Г. Опыт экспериментальной реализации алгоритмов поверхностно-синтаксического анализа // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог 2006, Бекасово, 31 мая – 4 июня 2006 г. — М.: Наука, 2006. — С. 51-56.

030-Kobozeva_Orlova:_Layout 1 18.05.2008 15:14 Page 199

ОДНОКЛЕТОЧНЫЕ ОРГАНИЗМЫ ОБЩЕНИЯ ПОД МИКРОСКОПОМ: НЕМЕЦКАЯ ЧАСТИЦА *JA* В СОПОСТАВЛЕНИИ С ЕЕ ПЕРЕВОДНЫМИ ЭКВИВАЛЕНТАМИ *ВЕДЬ* И *ЖЕ*

UNICELLULAR ORGANISMS OF COMMUNICATION UNDER A MICROSCOPE: GERMAN PARTICLE *JA* VERSUS ITS RUSSIAN TRANSLATION EQUIVALENTS *VED'* AND *ŽE*

Кобозева И.М. *(kobozeva@list.ru),* ***Орлова С.В.*** *(svetlachok-star@yandex.ru)*
Московский государственный университет имени М.В. Ломоносова

На материале параллельных фрагментов немецких и русских текстов проводится контрастивный анализ частицы ja и ее переводных эквивалентов ведь и же в контексте констативов и выявляются семантические и семантико-синтаксические различия в функционировании этих частиц, не зафиксированные в словарях и грамматиках.

Общей чертой русской и немецкой диалогической речи является обилие частиц, передающих информацию субъективно-модального (интенционального) и метатекстового характера. Активное изучение таких дискурсивных слов (далее ДС) в обоих языках, начавшееся в 80-е годы прошлого века, привело к появлению детальных семантических описаний основных русских и немецких ДС (см. обзор [4]). Это создает основу для разработки более точных правил перевода ДС с одного языка на другой, чем те, которые зафиксированы в двуязычных словарях и учебниках для иностранцев. Однако разработке таких правил должен предшествовать контрастивный анализ ДС обоих языков, выявляющий сходства и различия в их употреблении. Опыт именно такого предварительного анализа представляет собой проведенное нами сопоставление поведения немецкого ДС *ja* в тех его функциях, в которых его переводными эквивалентами считаются *же* и *ведь,* с поведением этих русских ДС.

Исходными данными для анализа служили: 1) переводные эквиваленты ДС *ja*, приводимые в немецко-русских словарях; 2) описания значения и употребления *ja*, *ведь* и *же* в семантических словарях служебных слов и специальных исследованиях; 3) выборка, содержащая около 200 фрагментов из немецких художественных текстов, в которых употреблено ДС ja, и их переводов на русский язык, выполненных профессиональными переводчиками высокого класса[1].

1. Ведь и же как эквиваленты ja по данным немецко-русских словарей

Как и большинство ДС, *ja* в немецком языке имеет несколько синтаксических и семантических функций, а в рамках последних представлено несколькими частными значениями. Одна из функций *ja* — функция утвердительной частицы-реплики, эквивалентной русской частице-реплике *Да*. Помимо этого *ja* функционирует и как «оттеночная частица» (Abtönungspartikel [8]), или, если использовать, возможно, менее удачный, но более привычный термин — модальная частица, имеющая ряд частных значений, передаваемых русскими ДС *обязательно*, *конечно*, *значит*, *просто*, и нек. др. Мы рассмотрим здесь лишь те употребления *ja*, в которых ее стандартными переводными эквивалентами считаются ДС *ведь* и/или *же,* как показывают следующие фрагменты словарных статей:

[1] В выборку вошли примеры из произведений Э.М. Ремарка и Кафки на немецком языке, опубликованные на сайтах « http://artefact.lib.ru и www.gutenberg.org, и их переводы, выполненные Шрайбером и Аптом и опубликованные на сайте http://lib.aldebaran.ru. Обращение к данным профессионального перевода мотивировано следующими соображениями: хороший переводчик старается максимально точно передать смысл оригинального текста и сделать это, не нарушая семантических и стилистических норм языка перевода. Поэтому, если переводчики в определенном типе контекстов регулярно предпочитают один из предлагаемых в словаре эквивалентов другим или же отказываются от словарных эквивалентов в пользу какого-то другого способа перевода, то это можно рассматривать как показатель того, что в данных контекстных условиях зафиксированная словарем переводная эквивалентность отсутствует и в них проявляют себя тонкие языковые различия между коррелятивными единицами разных языков или квазисинонимичными единицами языка перевода.

030-Kobozeva_Orlova:_Layout 1 18.05.2008 15:14 Page 200

(1) *ja* (…) 4. Zur Bekräftigung einer Ausssage (‘Для усиления высказывания’) synon. *doch*; *ведь, же*
Ich sagte ja schon! Ведь я уже сказал!
Du kennst ihn ja. Ведь ты его знаешь. Ты же его знаешь. (…) [9]
(2) *ja* (…) II. cj 1. ведь, же
Wir wollen zu Fuß gehen, es ist ja nicht weit bis dorthin. Мы пойдем пешком, туда ведь недалеко.
Warum sind die Kinder so vergnügt? – Es schneit ja! Чему так радуются дети? — Ведь идет снег.
Gedulde dich etwas, es kann ja nicht immer so bleiben. Потерпи немного, так ведь будет не всегда. [1]

Что естественно для переводных словарей, в словарных статьях, во-первых, не эксплицируются контекстные условия, в которых надо выбирать указанные в данном пункте переводные эквиваленты (они должны извлекаться пользователем из типовых примеров), а во-вторых, не указываются критерии выбора между двумя переводными эквивалентами.

Наша задача состояла в том, чтобы попытаться дать ответы на эти вопросы, опираясь на подробное описание функционирования частицы *ja* в словаре Г. Хельбига [8], а также на описание семантики русских *же* и *ведь* в словарях [6]и[7], и в особенности на их контекстно-семантический анализ в работе К. Бонно и С. В. Кодзасова [2].

2. Инвариантные значения ДС ja и ее переводных эквивалентов

Г. Хельбиг подводит все (не только модальные) употребления *ja* под одно общее значение, которое формулирует так: «Общее значение: состоит в закреплении, фиксации (Festlegung) Говорящим утверждаемого содержания (положение вещей мыслится как правильное или соответствующее действительности (zutreffend), предполагается возможность наличия соответствующих знаний у Слушающего), что наиболее четко проявляется в функции ответной частицы-реплики, но заметно также и в других функциях, когда частица соединяется с другими компонентами (напр., согласия, неожиданности) или когда они ее «перекрывают» [8].

Инвариантное значение *ведь*, согласно [Бонно, Кодзасов 1998], таково: «*Ведь* указывает на то, что вводимая информация, будучи адекватной, является одновременно релевантной для правильной интерпретации ситуации адресатом речи. Гарантом адекватности является действительность, гарантом релевантности – говорящий». И. Б. Левонтина [5] основной семантическим компонент в значении *ведь* представляет в виде ‘Я знаю и считаю, что это нужно принять во внимание’.

Из-за различия используемых авторами семантических метаязыков не так-то просто увидеть, какие семантические свойства являются у *ja* и *ведь* общими, а какие различными. Несомненно, общим является то, что обе частицы не предполагают обязательного наличия вводимой информации у Слушающего (далее С), а предполагают это только как возможность. Общим является также и то, что и *ja* и *ведь* маркируют вводимое пропозициональное содержание как в некотором смысле адекватное, правильное. Но совпадают ли гаранты адекватности? В другой формулировке, тождественны ли пропозициональные установки Говорящего (далее Г), имплицированные в семантике ДС? *Ведь* в *ведь P* имплицирует установку ‘Г знает, что Р имеет место’. Именно поэтому *ведь* не употребляется в императивах со значением прямого побуждения к действию [3][2]: в предусловие побуждения к действию (*Сделай P!*) входит установка ‘Г полагает, что Р не имеет места’, противоречащая установке *ведь*. ДС *ja* свободно употребляется в императивах, ср. (3):
Komm ***ja*** *nicht zu spät! — Смотри только приходи без опоздания! /* ****Ведь*** *приходи без опоздания!*
а это значит, что немецкое ДС имплицирует не установку знания, а установку оценочного мнения: ‘Г считает Р правильным (независимо от того имеет ли Р место в действительности)’.

Инвариантное значение ДС *же* К. Бонно и С. В. Кодзасов, перефразируя Д. Пайара [11], видят в том, что оно «маркирует сохранение точки слежения в сфере введенного в предтексте когнитивного объекта» [2, 406]. В такой формулировке значения *же* не легко увидеть, что общего у него со значением *ja*. Однако общность обнаруживается на уровне частных значений (граней). Так, и *ja* и *же*, равно как и *ведь* используются при введении информации в качестве аргумента за или против некоторого «тезиса». Заметим, что в этом типе контекстов упомянутые авторы связывают *же* не с гарантиями адекватности (в отличие от *ведь*), а только с гарантиями релевантности. Вместе с тем, в том же типе контекстов проявляется и отличие *же* как от *ja,* так и от *ведь*, вытекающее из специфики общего значения *же*, которое мы рассмотрим ниже.

Итак, соотношение инвариантных значений сопоставляемых ДС предсказывает существование трех типов контекстов (функций): 1) тех, в которых *ja* семантически эквивалентно *ведь*; 2) тех, в которых *ja* эквивалентно *же* (и сочетанию *ведь … же*); 3) тех, в которых ни *ведь*, ни *же* не являются эквивалентами ДС *ja*.

[2] Возможность *ведь* в риторических императивах типа «Ведь ты не забудь, когда это было!» объясняется в [5]

030-Kobozeva_Orlova:_Layout 1 18.05.2008 15:14 Page 201

3. Сопоставление **ja** *с* **ведь** *и* **же** *в констативных высказываниях*[3]

Согласно Г. Хельбигу, в этих контекстах *ja* выступает в значении ja_1: сигнализирует о том, что указанное положение вещей известно Г и С (=как мы оба знаем) или что оно очевидно, не подлежит сомнению; частица отсылает к общему знанию, предполагает согласие (основа коммуникации) или призывает к нему[4]. Г предполагает, что положение вещей известно, но хотел бы удостовериться, что оно является текущим (словно напоминает о нем)

3.1. **Ja, ведь** *и* **же** *в контексте аргумента, обоснования, объяснения*

Функция введения аргумента — общая для всех трех ДС. Различия между ними связаны прежде всего с коммуникативным (актуализационным) статусом вводимого сообщения.

По Хельбигу, ja_1 отсылает к общему знанию, то есть Г предполагает, что сообщаемое известно С, но, желая удостовериться, что оно находится в текущем сознании С, как бы напоминает о нем. Таким образом, *ja* приписывает вводимому аргументу статус «известного», но не «данного».

При помощи *ведь* Г может вводить как известную, так и неизвестную С информацию. Но даже когда Г знает, что С хорошо осведомлен о сообщаемом, *ведь* (как и *ja*) вводит информацию как «новое»: актуализирует некоторое знание у С, не будучи уверенным, что оно находится в его текущем сознании, как в (4):

(4) *Пора ложиться. [**Ведь** тебе завтра рано вставать.]* (пример из [2])

В отличие от этого, используя *же*, Г исходит из предположения, что вводимая в качестве аргумента информация, даже если она и не входила в фонд знаний С, уже активирована в его сознании (хотя бы с модальностью возможности), так как находится «в границах введенного в предтексте когнитивного объекта». Так, употребление *же* в (5), на наш взгляд, не обязательно связано с предположением Г, что приглашаемым на дачу известно о наличии там хороших леса и реки, но свидетельствует о том, что, по мнению Г, поездка на дачу подразумевает, включает в себя походы в лес и на реку:

(5) *Приезжайте к нам на дачу, обязательно приезжайте. [Там **же** такой лес, такая река!]* (пример из [12])

Таким образом, *же* приписывает вводимой информации статус «данного» (см. об этом в [7, 65]). А это, в свою очередь, придает высказыванию качество, которое в [2] названо «риторической активностью»: вводя аргумент при помощи *же,* Г «хочет залучить партнера в свидетели» того, что при наличии ситуации, которая приводится в качестве аргумента, ничего иного, чем принять вводимый тезис, не остается.

В случае аргумента «против», риторическая активность предложения с *же*, выступает в виде имплицитного упрека адресату, так как тот неадекватно ведет (вел, собирается вести) себя в некоторой ситуации, игнорируя те обстоятельства, которые, по мнению Г, адресат не мог не принимать в расчет, ср. (6):

(6) – *Почему ты не спросил ее вчера? [Она **же** была там.] И не говори, что ты ее не видел – она сидела прямо перед тобой!* (пример из [11])

Естественно, что *ведь*, не приписывающее сообщаемому статус «данного», лишено тех негативных импликаций, которые свойственны *же* в этом типе контекстов, ср. (9):

(7) – *Почему ты не спросил ее вчера? [Она **ведь** там тоже была.] Ты что ли ее не видел?* (пример из [11])

Итак, поскольку *ja* в аргументах по своим актуализационным свойствам (новизна / данность сообщаемого) аналогично *ведь*, а не *же*, то выбор *же* при переводе *ja* в составе аргумента должен быть мотивирован присутствием в контексте каких-либо причин для критики С за игнорирование очевидного и / или заинтересованностью Г в повышении риторической активности своего высказывания, иными словами в оказании воздействия на С в собственных интересах. Проанализировав наш материал, мы обнаружили, что переводчики интуитивно улавливают семантическую специфику русских ДС и выбирают *же* именно в таких случаях, см. типичный пример (8):

(8) *Später aber mußte man sie mit Gewalt zurückhalten, und wenn sie dann rief: «Laßt mich doch zu Gregor, [er ist **ja** mein unglücklicher Sohn!]. Begreift ihr es denn nicht, daß ich zu ihm muß?»…*

*Позднее удерживать ее приходилось уже силой, и когда она кричала: «Пустите меня к Грегору, [это **же** мой несчастный сын!] Неужели вы не понимаете, что я должна пойти к нему?»…*

В прочих случаях аргумент вводится при помощи нейтрального *ведь*, как в (9):

[3] По причине ограничений объема публикации мы оставляем в стороне другие иллокутивные типы контекстов в которых частице *ja* соответствуют *ведь* и / или *же*.

[4] Этим *ja* отличается от *doch*, при помощи которого Г вводит правильную с его точки зрения информацию в противовес высказанному или предполагаемому противоположному мнению С. Для русских *ведь* и *же* компонент ‘согласия’ нерелевантен.

030-Kobozeva_Orlova:_Layout 1 18.05.2008 15:14 Page 202

(9) *«Trinken wir etwas», sagte ich. [«Es ist **ja** noch nichts verloren.»]*
*Выпьем немного, – сказал я. – [**Ведь** еще ничто не потеряно.]*

Не удивительно при этом, что когда Г адресует аргумент себе самому, или когда вводимая информация — это не аргумент за или против тех или иных действий С, а обоснование или объяснение некоторого утверждения Г, непосредственно не затрагивающего его интересы, то *ja* или соответствует *ведь* как в (10), (11) или вообще опускается при переводе, как в (12):

(10) *Ich war etwas verlegen und wußte nicht recht, wie ich ein Gespräch anfangen sollte.[Ich kannte das Mädchen **ja** überhaupt nicht…]*
*Я был несколько смущен и не знал, с чего начинать разговор. [**Ведь** я вообще не знал эту девушку…]*

(11) *Ich war etwas erstaunt, ihn so plötzlich weich zu sehen, und vermutete, daß ihm das flinke schwarze Luder, das er zuletzt bei sich gehabt hatte, bereits auf die Nerven ging. [Ärger macht **ja** die Leute leichter sentimental als Liebe.]*
*Я слегка удивился, увидев его вдруг таким размякшим, и предположил, что шустрая чернявая бабенка, которая приходила с ним в последний раз, уже начала действовать ему на нервы. [**Ведь** люди становятся сентиментальными скорее от огорчения, нежели от любви].*

(12) *«Fahr hin! ... Sag ihm, daß das mit Gottfried fertig ist. Habe früher Bescheid gewußt als ihr! [Siehst es **ja**!]».*
– Поезжай туда! ... Скажи ему, что за Готтфрида я расквитался. Я знал об этом раньше вас! [Сам видишь], что я ранен!

И только когда повествователь вводит в качестве обоснования сказанного информацию, затрагивающую интересы персонажа, переводчик может как бы начать говорить от лица персонажа, используя форму свободного косвенного дискурса, и перевести *ja* при помощи *же*, как в (13):

*(13) In der ersten Zeit stellte sich Gregor bei der Ankunft der Schwester in derartige besonders bezeichnende Winkel, um ihr durch diese Stellung gewissermaßen einen Vorwurf zu machen. Aber er hätte wohl wochenlang dort bleiben können, ohne daß sich die Schwester gebessert hätte; [sie sah **ja** den Schmutz genau so wie er], aber sie hatte sich eben entschlossen, ihn zu lassen.*

*Первое время при появлении сестры Грегор забивался в особенно запущенные углы, как бы упрекая ее таким выбором места. Но если бы он даже стоял там неделями, сестра все равно не исправилась бы; [она **же** видела грязь ничуть не хуже, чем он], она просто решила оставить ее.*

Здесь *же* в авторской речи есть отображение молчаливо посылаемого Грегором упрека сестре: «Почему ты не убираешься в комнате? Ты **же** видишь грязь не хуже, чем я».

3.2. Ja, ведь и же в контесте реплик-реакций

3.2.1. Корректирующие реакции.

Ja, ведь, и *же* употребляется в составе реплик, в которых Г вводит информацию, показывающую, что в инициативной реплике содержится та или иная ошибка. Это может быть констатация нарушения пресуппозиций, как в (14):

(14) *«Die Dame, die Sie immer verstecken«, sagte Frau Zalewski, »brauchen Sie nicht zu verstecken. Sie kann ruhig offen zu Ihnen kommen. Sie gefällt mir...»*
*«[Sie haben sie **ja** noch gar nicht gesehen]», erwiderte ich.*
– Даму, которую вы всегда прячете от нас, – сказала фрау Залевски, – можете не прятать. Пусть приходит к нам совершенно открыто. Она мне нравится.
*Но [вы **ведь** ее не видели], – возразил я.*

где вторая реплика констатирует нарушение пресуппозиции оценочного суждения «Она мне нравится» (предусловием всякой оценки является ознакомление с объектом оценки). Это может быть указание на несоблюдение партнером по коммуникации условий успешности речевого акта, как в (15):

(15) *«Herr Samsa», rief nun der Prokurist mit erhobener Stimme, «was ist denn los? Sie verbarrikadieren sich da in Ihrem Zimmer, antworten bloß mit ja und nein, machen Ihren Eltern schwere, unnötige Sorgen und versäumen — dies nur nebenbei erwähnt — Ihre geschäftlichen Pflichten in einer eigentlich unerhörten Weise.»(...)*
*«Aber Herr Prokurist», rief Gregor außer sich und vergaß in der Aufregung alles andere, «[ich mache **ja** sofort, augenblicklich auf]».*
– Господин Замза, – воскликнул управляющий, теперь уж повысив голос, – в чем дело? Вы заперлись в своей комнате, отвечаете только «да» и «нет», доставляете своим родителям тяжелые, ненужные волнения и уклоняетесь – упомяну об этом лишь вскользь – от исполнения своих служебных обязанностей поистине неслыханным образом.
*– Но, господин управляющий, – теряя самообладание, воскликнул Грегор и от волнения забыл обо всем другом, – [я **же** немедленно, сию минуту открою].*

030-Kobozeva_Orlova:_Layout 1 18.05.2008 15:14 Page 203

где сообщение Грегора в ответ на обвинения управляющего, пытающегося таким способом заставить его открыть дверь, имеет целью дать собеседнику знать, что тот ошибается, предполагая отсутствие у него намерения это сделать.

Как и в аргументативном контексте, выбор *же* сопряжен с контекстными условиями, способными оправдать повышение «риторической активности», экспрессивности реплики. Кроме этого семантико-прагматического фактора, как нам удалось обнаружить, на выбор эквивалента для *ja* в данном типе контекстов влияют и другие, семантико-синтаксические факторы.

Так, существует ограничение на употребление *ведь*, отсутствующее у *ja* и *же*: *ведь* не может вводить реплику, констатирующую нарушение условия успешности речевого акта, если пропозиция второй реплики во всем, кроме отрицания и модальности, повторяет пропозицию первой. Так, невозможна замена *да* — основного маркера реплик, сообщающих о нарушении условий успешности — на *ведь* в (17):

(16) *«Du mußt nicht weinen», sagte ich. ... «[Ich weine **ja** gar nicht]», erwiderte sie ...*
*Ты не должна плакать, – сказал я. ...– [**Да/*Ведь** я **и** не плачу], – проговорила она…*

Замена *да* на *же* в принципе возможна (ср. *Я же не плачу*), но она неизбежно повышает риторическую активность реплики. Таким образом, в указанной разновидности контекстов нейтральным эквивалентом *ja* выступает *да*, что не отмечено в словарях.

Сопоставление трех ДС в контексте корректирующей реплики показало еще одно семантико-синтаксическое различие, которое ранее не отмечалось. *Же*, как и *ja* при введении такой реплики, являются самодостаточными ДС в том смысле, что не требуют (хотя и не исключают) появления дополнительных маркеров связи второй реплики с первой. Так, переводом второй реплики в (14) могла бы быть реплика (14') (при соответствующем повышении «градуса» экспрессивности):

(14') *– Вы **же** ее не видели, – возразил я.*

В отличие от этого, *ведь* в корректирующей реплике-реакции практически не встречается без инициальных частиц *а, но*, или *да,* сигнализирующих о «противительном» отношении второй реплики к первой. Так, в (17) первая реплика с *ведь*, указывающая на нарушение пресуппозиций благодарности (отсутствие того, за что благодарят), равно как и соответствующая ей реплика с *ja* начинается с союза-частицы, и только при повторе незаконченной ранее реплики *ведь* выступает без поддержки такого ДС:

(17) *«Schatzi!» Mit einem Satz hing sie dem Bäcker am Hals.*
*«Aber [das ist **ja** noch gar nicht...]» Er versuchte sich loszumachen und Erklärungen abzugeben. (...)*
*«[Wir sind **ja** noch gar nicht soweit]», prustete er.*
– Сокровище ты мое! – Она подпрыгнула и повисла на шее у булочника.
*– [**Но ведь** мы еще…] – Он пытался высвободиться из ее объятий и объяснить ей положение дел. (...)*
*– [**Ведь** мы еще не договорились], – сказал он, отдуваясь.*

Следует отметить, что корректирующая реплика в любом случае нарушает максиму согласия (одну из максим, обеспечивающих бесконфликтное общение в соответствии с принципом вежливости, см. [10]), а присутствие в ней явного показателя «противительности» делает ее более конфликтной. Поэтому, вводя противительный союз там, где его не было в оригинале, как это сделано в (14), переводчик изменяет прагматические характеристики реплики в сторону ее большей конфликтности. Таким образом, при переводе *ja* в составе корректирующей реплики без противительного союза возникает коллизия: надо или отказаться от перевода *ja* как *ведь*, или вводить противительный союз, меняя интеракционную характеристику высказывания. Последнее делается в тех случаях, когда взаимоотношения персонажей в данной ситуации допускают некоторое повышение конфликтности реплики, как в (14), где герой обращается к квартирной хозяйке, и тем более в (18), где полицейский адресуется к возмутителю спокойствия:

(18) *Ein Schupohelm blitzte. «Was ist hier los?» ... «Nichts», sagte ich.... «[Sie bluten **ja**.]»*
Сверкнула каска полицейского. – Что здесь случилось?...– Ничего, – сказал я....
*– [**Но ведь** вы в крови].*

В тех же случаях, когда нет оснований ни для повышения конфликтности корректирующей реплики, ни для повышения ее риторической активности (что позволило бы использовать *же* вместо *ведь* и не вводить конфликтогенное ДС), можно наблюдать, как переводчик отказывается от перевода *ja* при помощи *ведь* или *же*, прибегая к другим средствам, например, к использованию эмфатической препозиции ремы, как в (19):

(19) *«Deine Sachen nehmen wir alle mit», sagte ich. «Du sollst hier nichts entbehren. Sogar einen Teewagen schaffen wir uns an....»*
*«[Wir haben **ja** einen], Liebling. ...»*
– Все твои вещи перевезем сюда, – сказал я. – Чтобы у тебя здесь было все. Даже заведем чайный столик на колесах....
– [Есть у нас такой столик], милый....

030-Kobozeva_Orlova:_Layout 1 18.05.2008 15:14 Page 204

3.2.2. Ответы на вопрос.

Как известно, одной из основных функций *ja* является выражение положительного ответа на общей вопрос, и в этой функции ее эквивалентом выступает русское *да*. Однако, *ja* в ответах на вопрос может выступать и в модальной функции. Во-первых, вопреки утверждению Г. Хельбига [8], модальное *ja* встречается в прямых ответах на некоторые виды частных вопросов, а именно в вопросах, ответ на которые представляет собой полную пропозицию, что имеет место при вопросах о причине и цели, как в примере из [1], повторяемом в (20):

(20) *Warum sind die Kinder so vergnügt? – Es schneit ja!*

В [1] предложен перевод (20'):

(20') *Чему так радуются дети? — Ведь идет снег.*

На наш взгляд, перевод ответа не вполне естественный по причине отсутствия перед *ведь* какой-либо инициальной частицы, которая, как показал беглый просмотр примеров из НКРЯ, всегда сопровождает прямые ответы с *ведь*. По-видимому, это русское ДС не может быть единственным маркером смысловой связи высказываний, принадлежащих разным субъектам, если только высказывание второго не является прямым продолжением высказывания первого, как бы договариванием за него. *Же* в данном типе контекстов может употребляться, как и в контекстах типа 1.3.2, без поддержки инициальных частиц, ср. (20'') *Чему так радуются дети? — Им же подарки подарили.*

Все 3 ДС в подобных случаях не просто вводят запрашиваемую С информацию как несомненно достоверную, но и выражают отношение эпистемической обусловленности: 'если бы С знал это, у него не возникло бы вопроса'. Разница между *ведь* и *же* опять-таки состоит в большей риторической активности *же*, обусловленной ее семантическим инвариантом.

Второй тип ответов, в которых нам встретилось модальное *ja*, — это косвенные ответы на общий вопрос, сообщающие несомненно достоверную информацию, которая делает очевидным положительный или отрицательный ответ на него. Ни *ведь*, ни *же* в таких случаях использоваться не могут, и как показал анализ материала, переводчики просто игнорируют *ja*, как в (20):

(20) *«Alles schon vorbereitet, Lilly?» fragte ich. Sie nickte. [«Die Aussteuer hatte ich **ja** schon lange»].*

– Все уже приготовлено, Лилли? – спросил я. Она кивнула: – [Приданым я запаслась давно]

Ср. *—?Ведь приданым я запаслась давно; — ?Приданым же я запаслась давно.*

4. Выводы

Сопоставление немецкого модального ДС *ja* с русскими *же* и *ведь* в контексте констативов показало, что все три ДС могут использоваться при введении аргумента, обоснования или объяснения, при коррекции логических и коммуникативных ошибок партнера, а также в прямых ответах на вопросы о причине и цели. *Ja*, в отличие от *ведь* и *же*, встречается и в ответных репликах, имплицирующих ответ «да» или «нет» на общий вопрос. При этом *ведь* и *же*, в отличиие от *ja*, специфицируют степень риторической активности Г, обусловленную либо его прямой заинтересованностью в предмете речи либо критическим настроем по отношению к С: *ведь* — маркер нормы по этому параметру, а *же* — маркер превышения нормы, что делает выбор *ведь* и *же* при переводе небезразличными к таким коммуникативно-прагматическим параметрам контекста, как отношение обсуждаемой темы к интересам говорящего, а также межличностные отношения между коммуникантами. Помимо этого выявлены семантико-синтаксическое ограничения на употребление *ведь* в контексте корректирующих и ответных реплик: 1) *ведь* требует присутствия в реплике инициального противительного ДС (*но*, *а* , *да*); 2) *ведь* требует различия диктальных компонентов пропозиционального содержания инициативной и реактивной реплик. Выполнение первого требования при переводе повышает конфликтность переводной реплики по сравнению в оригинальной, что может нарушить смысловую эквивалентность высказываний в интеракционном аспекте. Такая ситуация разрешается либо путем поиска иных способов перевода, либо путем игнорирования вклада *ja* в семантику исходного высказывания.

Результаты проведенного контрастивного анализа могут найти применение в практике преподавания немецкого и русского языков как иностранных, а также использоваться для совершенствования переводных словарей.

030-Kobozeva_Orlova:_Layout 1 18.05.2008 15:14 Page 205

Список литературы

1. Большой немецко-русский словарь в 3 томах. Под ред. О.И. Москальской. Изд. 5-е, стереотип. М., 2001.

2. Бонно К., Кодзасов С.В. Семантическое варьирование дискурсивных слов и его влияние на линеаризацию и интонирование (на примере частиц же и ведь) // Дискурсивные слова русского языка: опыт контекстно-семантического описания. Под ред. К. Киселевой и Д. Пайара. М., 1998, 382-443.

3. Кобозева И. М. Русские модальные частицы и их согласование с иллокутивной функцией высказывания. // Linguistische Arbeitsberichte, B. 70, 1988, 38-47.

4. Кобозева И. М. Проблемы описания частиц в исследованиях 80-х годов // Прагматика и семантика. М.: ИНИОН АН СССР, 1991, 147-176.

5. Левонтина И. Об одной загадке частицы ВЕДЬ //В сб.: Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2005». М., 2005.

6. Объяснительный словарь русского языка. Под ред. В. В. Морковкина. М., 2003.

7. Шимчук Э., Щур М. Словарь русских частиц. Berlin, 1999.

8. Helbig G. Lexikon deutscher Partikeln. Leipzig: VEB Verlag Enzyklopädie, 1988.

9. Langenscheidts Großwörterbuch Deutsch-Russisch, 1997.

10. Leech, G. Principles of Pragmatics. Longman, London, 1983

11. Paillard D. ŽE ou la sortie impossible // Les particules enonciative en russe contemporain (2). Paris, 1987.

12. Vasilyeva A. N. Particles in colloquial Russian. M., 1972.

031-Kodzasov_S:_Layout 1 18.05.2008 15:39 Page 206

БАЗА ДАННЫХ «ИНТОНАЦИЯ РУССКИХ ИНФОРМАЦИОННЫХ ТЕКСТОВ»[1]

DATABASE «INTONATION OF RUSSIAN INFORMATIONAL TEXTS»

Кодзасов С.В. *(sankod@philol.msu.ru),* ***Архипов А.В.*** *(arxipov@philol.msu.ru),*
Захаров Л.М. *(leon@philol.msu.ru),* ***Кривнова О.Ф.*** *(okri@philol.msu.ru)*
Московский государственный университет имени М.В. Ломоносова

В докладе сообщается о результатах работы по созданию БД «Интонация русских информационных текстов». Этот проект продолжает многолетнее исследование, направленное на создание понятийного аппарата и компьютерного инструментария для всестороннего описания функций и средств русской интонации на основе результатов анализа представительного корпуса русских текстов (диалогических и повествовательных). В работе обсуждаются результаты начального этапа второго нового цикла наших корпусных исследований.

1. Описание проекта

Проект связан с быстро развивающейся областью отечественной лингвистики, которую можно назвать дискурсивной грамматикой русского языка. В ее задачи входит описание дискурсивной лексики, дискурсивного синтаксиса и дискурсивной интонации. До сих пор корпусные методы применялись изучалась в основном для исследования интонациия диалогического дискурса. Результаты исследований разработок авторского коллектива в этой области сообщались на прошлых конференциях «Диалог» (см. список публикаций).

В осуществляемом ныне проекте впервые ставится задача создания компьютерной речевой базы для изучения интонации информационного и повествовательного дискурса. При этом просодия текстов исследуется демонстрируется и описывается в тесной связи с их иллокутивным и модальным содержанием, а также с их фазово-синтаксической структурой. Особое внимание уделяется просодическим техникам маркированияки жанровых и регистровых характеристик текстов и их компонентов. Это позволит уточнить системы семантических, прагматических и стилистических дескрипторов, которые используются в текст-лингвистике в настоящее время.

Результаты проекта могут использоваться как при дальнейших лингвистических исследованиях грамматики текста, так и в прагматических целях (оптимизация учебных и инструктивных текстов, рекомендации по использованию произносительных стилей в средствах массовой информации). В число потенциальных пользователей попадают не только лингвисты, но также дикторы и составители текстов в средствах массовой информации.

Проект предполагает создание трех представительных корпусов звучащих текстов для трех базовых областей дискурса: информационные и инструктивные тексты (2007 г.), нарративные тексты разных жанров (современные литературные тексты, фольклорные тексты) (2008 г.), спонтанные нарративы (2009 г.). Характерные фрагменты для каждого типа дискурса будут введены в компьютерную базу данных. Каждый файл снабжается краткой прагматико-семантической дескрипцией и просодической транскрипцией, включающей все просодические характеристики: тональные акценты и интегральные уровни тона, выделительные акценты и громкости, долготы и темпы, фонации и просодические тембры. План работ по годам соответствует трем указанным областям дискурса.

До сих пор лингвистами с помощью корпусных методов изучалась в основном интонация диалогического дискурса. В данном проекте впервые была поставлена задача создания компьютерной базы для изучения интонации информационного и повествовательного (литературного и бытового) дискурса.

За отчетный период в целях создания корпусного материала были исследованы просодические средства оформления текстов информационных программ центрального телевидения. В качестве материала использовались записи новостных программ трех наиболее популярных телеканалов: НТВ («Сегодня»), ОРТ («Время») и РТР («Вести»), которые стилистически различны.

[1] Работа поддержана грантом РГНФ № 07-04-12160в.

031-Kodzasov_S:_Layout 1 18.05.2008 15:39 Page 207

Исследовались акцентная структура текстов (размещение семантически нагруженных выделительных акцентов), тональная структура компонентов сообщения и средства экспрессивного подчеркивания. Производился анализ функциональной структуры текста (фазовые и модальные характеристики его компонентов), что позволяет выделить средства семантического структурирования сообщения и риторического воздействия текста на слушающего.

В результате анализа было отобрано около 300 текстовых фрагментов (информационных блоков) общим временем звучания около 90 минут. Они введены в базу данных в виде звуковых файлов, снабженных несколькими типами аннотаций: простейшая функциональная (тематическая) индексация, орфографическая запись, просодическая транскрипция текстового фрагмента.

Каждый файл снабжен краткой семантико-прагматической дескрипцией и детальной просодической дескрипцией на основе разработанной авторами ранее системы интонационного описания (см. список публикаций). Эта информация представлена в виде базы данных в формате Microsoft Access. Программа позволяет прослушивать файл, видеть его осциллограмму и интонограмму и анализировать семантико-просодическую дескрипцию.

2. Пример записи информационного блока в базе данных («Новости» на НТВ)

Тема блока и темы фрагментов	**Орфографическая запись фрагмента**	**Просодия фрагмента**
БЛОК 1. ПОЖАР В ДОМЕ ПРЕСТАРЕЛЫХ		
1.1. Развитие чрезвычайной ситуации, число жертв	Число погибших при пожаре в доме престарелых под Тулой сегодня вновь возросло и достигло тридцати двух человек.	Число(\)(\) *погибших(\) при пожаре в доме *престарелых(/) | под *Тулой (\\) | сегодня *вновь (/) *возросло(\\) и *достигло(\) *тридцати(\) двух *человек (\).
1.2. Еще одна жертва	Уже после того, как основные поиски были Взавершены, на месте сгоревшего интерната кинологи обнаружили тело еще одной жертвы.	Уже после *того (/\), как основные(/) *поиски(/) были *завершены(/), || на *месте(/) сгоревшего(/) *интерната(/) кинологи(/) *обнаружили(\) *тело(\) *еще(\) одной *жертвы(\).
1.3. Причина большого числа жертв	В МЧС говорят, что масштабной трагедии можно было избежать. Медики сами пытались тушить огонь и долго не сообщали о возгорании по ноль один.	В *МЧС говорят(\), | что *масштабной(\) *трагедии(/В) можно было *избежать(\). || Медики(/) *сами пытались(\) тушить *огонь(//) | и долго не *сообщали(\\) о *возгорании(\) по ноль *один(\).
1.4. Последствия	Сегодня в Тульской области объявлен день траура. Тем временем в городах и поселках началась проверка пожарной безопасности в старых зданиях.	*Сегодня(\) в Тульской(\) области(\) | объявлен(/) день(/) *траура (\\). || Тем *временем(/) | в городах и *поселках(\\) | началась *проверка(/) пожарной *безопасности(/) в старых *зданиях(\\).
1.5. Текущее состояние 1	С места события передает наш специальный корреспондент Евгений Гуцал. — Все работы в доме престарелых сейчас завершаются. Режим чрезвычайной ситуации в Тульской области уже снят.	С места *события(\) передает наш *специальный(\)(\) *корреспондент(/) Евгений *Гуцал(\). Все *работы(/) в доме *престарелых(+) сейчас *завершаются(\). *Режим (/) *чрезвычайной *ситуации в *Тульской(\) *области(/) уже(/) *снят(\).
1.6. Число погибших и их похороны	Окончательное число погибших — тридцать два человека. Всех их похоронят за счет государства. У многих нет даже родственников.	Окончательное(/\) число *погибших (/) || — *тридцать(\) *два(\) человека. Всех(\) их похоронят(\) за счет *государства(\). У многих(/) нет даже *родственников (\).
1.7. Текущее состояние 2	Сегодня утром составлены списки: кто из пациентов сейчас в больнице, кого отправили в ближайшие интернаты. К месту трагедии съезжаются близкие пострадавших.	*Сегодня(/) *утром(/) *составлены(\)(\) списки: кто из *пациентов(\) *сейчас(\) в *больнице(/), *кого(/) отправили в *ближайшие(\) *интернаты(\)(\).|| К месту *трагедии(/) *съезжаются(\)(\) *близкие(\) пострадавших.

031-Kodzasov_S:_Layout 1 18.05.2008 15:39 Page 208

Тема блока и темы фрагментов	Орфографическая запись фрагмента	Просодия фрагмента
1.8. Подтверждение прежней версии причины события	Продолжают работать следователи. Официальной версией пока так и остается возгорание проводки.	*Продолжают(/) работать(/) *следователи(\).|| Официальной *версией(/) пока так и *остается(\) возгорание *проводки.
1.9. Последние данные с места события	О том, что найдено тело тридцать *второго постояльца, стало известно во время заседания специальной комиссии по расследованию ситуации.	О том, что найдено тело тридцать *второго(+) постояльца(\),|| стало известно(/) во время заседания специальной *комиссии(/) по расследованию(/) ситуации(\).
1.10. Дальнейшая судьба дома престарелых	Сотрудники МЧС заявили, что здание дома престарелых восстановлению не подлежит. В области сейчас решается вопрос о строительстве нового интерната.	Сотрудники *МЧС *заявили(/), что здание дома *престарелых (/) восстановлению(/) не *подлежит(/).|| В области сейчас решается *вопрос(/) | о строительстве || *нового(\) интерната.
1.11а. Сопутствующие гибели людей мероприятия	Похороны погибших состоятся завтра.	Похороны(/) *погибших(/) состоятся(\) *завтра(\).
1.11б. Объявление траура	Сегодня в Тульской области объявлен траур.	*Сегодня(/) в Тульской *области(/) | объявлен *траур(\).

3. Просодическая специфика информационных текстов (предварительные наблюдения)

Материал нашей базы данных обнаруживает немало отклонений просодии телепрограмм от типичных характеристик диалогического текста, что видно и по просодической разметке приведенного выше фрагмента. Укажем наиболее важные их них.

Прежде всего, обращает на себя внимание гораздо большая степень акцентированности текста в целом. Это характерно не только для новостных программ НТВ, но и для новостей на других каналах. Нередки случаи выделения громкостными или тональными акцентами всех слов предложения (своего рода «акцентное скандирование» текста). Другая особенность: наличие многочисленных случаев несовпадения выделительных и тональных акцентов, что для реплик диалога нехарактерно. Немало также случаев двойных тональных акцентов внутри слова. Все это говорит о том, что в текстах рассматриваемого вида используется особая акцентная подсистема, ориентированная на высокую степень риторического воздействия на слушателя.

Другой особенностью проанализированных текстов является довольно частое использование восходящего тона в качестве финального акцента неконечных предложений внутри фрагмента сообщения. Риторическая функция этой техники очевидна: диктор указывает на незавершенность текста, мобилизуя внимание слушающего на его продолжение.

Мы приводим в Приложении расшифровку знаков используемой в данной работе просодической транскрипции, поскольку она не общепринята и может вызвать затруднения у читателя. Эта же система использовалась нами при создании баз данных по интонации русских диалогических реплик (см. список литературы).

Приложение. Знаки интонационной транскрипции

Знаки для тональных акцентов

\ - нисходящий нейтральный (4-7 полутонов)
\’ – нисходящий малый (3-4 полутона)
\” – нисходящий сверх-малый (менее 3 полутонов)
\\ - нисходящий увеличенного интервала
\в – нисходящий в высоком регистре
\н – нисходящий в низком регистре
\с – нисходящее движение происходит на начальном согласном слога
\~ - нисходящий тон растянут на слово
/ - восходящий нейтральный
/’ – восходящий малый
/” – восходящий сверх-малый

031-Kodzasov_S:_Layout 1 18.05.2008 15:39 Page 209

// - восходящий увелич. интервала
/в – восходящий в высоком регистре
/н – восходящий в низком регистре
/с – восходящее движение происходит на начальном согласном слога
/~ - восходящий тон растянут на слово
/-\ - восходящий – ровный – нисходящий
/\ - восходяще-нисходящий тон внутри гласного
/с\г – восходящий тон на согласном и нисходящий на гласном
\/ - нисходяще-восходящий тон внутри гласного
/- - восходящий плюс ровный без падения
Знаки для нетональных акцентов
: - долготный акцент
* – громкостный акцент на подчеркнутом сегменте слова
Знаки для синтагменных просодий
инк – инклинация тона
дек – деклинация тона
ТИ – низкая громкость (тихо)
ГР – высокая громкость (громко)
Н – низкий регистр
В – высокий регистр
Б – быстрый темп
НПР – напряженная фонация
ПДХ – придыхательная фонация
ФЦТ – фальцетный регистр

Список литературы

1. Кодзасов С.В., Бонч-Осмоловская А.А., Захаров Л.М., Кобозева И.М., Кривнова О.Ф. База данных «Интонация русского диалога»: вопросительные реплики // Труды международной конференции Диалог 2005 г.

2. Кодзасов С.В., Архипов А.В., Бонч-Осмоловская А.А., Захаров Л.М., Кривнова О.Ф. База данных «Интонация русского диалога»: побудительные реплики. // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции Диалог 2006 г.

3. Кодзасов С.В., Архипов А.В., Захаров Л.М., Кривнова О.Ф. База данных «Интонация русского диалога»: реплики-сообщения. Компьютерная лингвистика и интеллектуальные технологии // Труды международной конференции Диалог 2007 г.

032-Kozhunova_korr:_Layout 1 18.05.2008 15:24 Page 210

КЛАССИФИКАЦИОННАЯ СХЕМА СЕМАНТИЧЕСКОГО СЛОВАРЯ СИСТЕМЫ МОНИТОРИНГА: ОПЫТ ПРИМЕНЕНИЯ В ПРОЦЕССЕ ОЦЕНКИ РЕЗУЛЬТАТИВНОСТИ НАУЧНОЙ ДЕЯТЕЛЬНОСТИ

CLASSIFICATION SCHEME OF THE SEMANTIC DICTIONARY OF THE MONITORING SYSTEM: TEST APPLICATION TO EVALUATION OF SCIENTIFIC WORK' PERFORMANCE

Кожунова О.С. *(kozhunovka@mail.ru, okozhunova@ipiran.ru)*
Институт проблем информатики РАН

Кратко описывается эксперимент по оценке результативности научной деятельности в Российской Академии наук, проведенный в первом полугодии 2007 года. Рассматриваются выявленные по итогам этого мероприятия проблемы, для решения которых предлагаются метод и соответствующий инструментарий, а именно: классификационный метод и семантический словарь с интегрированной в него классификационной схемой.

Введение

Проведенный в первом полугодии 2007 года эксперимент по оценке результативности научной деятельности в Российской Академии наук показал, что сегодня существенно изменилось значение данных мониторинга научных исследований и деятельности субъектов сферы науки. В 2006 и 2007 годах был разработан ряд нормативных документов, направленных на распределение значительной части научного бюджета с учетом значений индикаторов результативности научных исследований (Совместный приказ Министерства образования и науки РФ, Министерства здравоохранения и социального развития РФ и Российской академии наук № 273/745/68; Устав РАН, утвержденный постановлением Правительства РФ от 19 ноября 2007 г. № 785; Постановление Правительства РФ от 22.04.2006 года № 236) [1, 2]. Таким образом, в настоящее время данные мониторинга стали оказывать существенное влияние на бюджетный процесс. Поэтому системы мониторинга, анализа и оценки (далее – системы мониторинга) и определяемые с их помощью значения индикаторов используются в качестве основных критериев в процессе принятия решений, согласования оценок результативности и вынесения окончательных оценок научной деятельности экспертами.

На сегодняшний день накоплен отечественный и зарубежный опыт мониторинга, анализа и оценки результативности в сфере науки [3, 4, 5]. Анализ этого опыта позволяет утверждать, что изменение роли систем мониторинга и определяемых с их помощью значений индикаторов придает весьма актуальный характер задаче построения словаря терминов для описания и решения широкого спектра задач этой области.

Это мотивировано тем, что системы мониторинга должны стать, во-первых, депозитариями *сертифицированных информационных ресурсов*, так как их данные начинают использоваться в процессах бюджетирования и принятия решений. Во-вторых, процедуры обеспечения процессов бюджетирования и принятия решений, реализованные в системе мониторинга, должны быть *легатимизированы*. В-третьих, в целях обеспечения инвариантности системы именно в семантическом словаре должны быть локализованы формы представления знаний экспертов, результатов согласования категоризации и степени полноты их знаний об окружающей системе и взаимодействующей с ней институциональной среде, о ее целях и решаемых ею задачах.

В работе в связи с этим были рассмотрены следующие аспекты:

- Краткое описание эксперимента по оценке индивидуальных показателей результативности научной деятельности, проведенного в Российской Академии наук в первом полугодии 2007 года;
- Проблемы, выявленные по итогам проведения эксперимента, и предлагаемые в данной работе метод и инструмент их решения.

Эксперимент по оценке результативности научной деятельности

Совместным приказом Министерства образования и науки РФ, Министерства здравоохранения и социального развития РФ и Российской Академии наук № 273/745/68 от 03.11.2006 с целью оценки научной

032-Kozhunova_korr:_Layout 1 18.05.2008 15:24 Page 211

работы на наноуровне[1] (ученые и инженеры) был утвержден список индикаторов [1]. В Приказе предполагалось инициировать анализ и оценку деятельности отдельных работников Академии наук с июня 2007 года при помощи приведенных в тексте этого нормативного документа оценочных количественных и качественных индикаторов. Для индикаторов были приведены некоторые единицы измерения, на основании подсчета которых вычислялись промежуточные и итоговые баллы результативности деятельности отдельно взятого ученого (или инженера). Количеству набранных баллов в рамках бюджета определенного подразделения Академии наук выставлялся в соответствие некоторый денежный эквивалент, выплачиваемый сотрудникам в виде надбавки по итогам такого исследования их результативности.

Поскольку данные проведенного эксперимента недоступны на уровне всей Академии наук, то описать его особенности возможно только на примере отдельного учреждения. В предлагаемой работе кратко описываются результаты мониторинга, анализа и оценки деятельности ученых и инженеров Института Проблем Информатики РАН (ИПИ РАН, 20 отделов и 2 филиала), поскольку автор является одним из разработчиков инструментария мониторинга и оценки индивидуальных показателей результативности научной деятельности, проведению оценок, анализу списка утвержденных Приказом (ссылка) индикаторов и вычислению итоговых показателей и оценок результативности для присваивания баллов сотрудникам.

В работе [6] уже говорилось о задаче согласования понимания терминов с целью разрешения терминологических разногласий между экспертами и необходимости в инструменте для их устранения. Проведенный эксперимент подтвердил актуальность использования именно семантического словаря в качестве такого инструмента, поскольку точность, наглядность и адекватность определений терминов (в частности, значений отдельных индикаторов) играет решающую роль в вычислении показателей результативности ученых и итоговой оценке их деятельности.

Например, пункт 2.3 вышеупомянутого Приказа [1]: «Индивидуальный ПРНД научных работников, работающих по совместительству, умножается на коэффициент, равный отношению продолжительности рабочего времени совместителя в месяц к нормальной продолжительности рабочего времени штатного работника на аналогичной должности. При этом в расчет должны приниматься только те результаты, которые получены при работе в Институте и официально к ней отнесены (наличие наименования Института, как места выполнения работы, в публикациях, материалах конференций и иных результатах научной деятельности, учитываемых при расчете индивидуального ПРНД, или публикация результатов в изданиях Института). Для работников, поступивших на работу в Институт не ранее, чем за два года до года выплаты надбавок стимулирующего характера, при расчете индивидуального ПРНД учитываются их результаты, полученные по основному месту работы».

При подаче материалов для вычисления индивидуальных показателей результативности научной деятельности (сокращенно ПРНД) многие сотрудники-совместители столкнулись с проблемой многозначного толкования этого пункта. Члены Комиссии по оценке ПРНД также были вынуждены дополнительно согласовать свое понимание пункта о совместителях и о новых сотрудниках перед проведением оценки. К сожалению, параграфов, подобных примеру, в Приказе оказалось немало.

Кроме того, в Приказе не было приведено дефиниций результатов научной деятельности и соответствующих индикаторов, характеризующих достижение того или иного результата и степень достижения, а также какой-либо классификации результатов и индикаторов. Из данного нормативного документа следовало, что цитируемость автора, его научное руководство и участие в конференциях, например, есть суть результаты одного уровня и одной категории.

Все эти аспекты существенно усложняли проведение мониторинга, анализа и оценки деятельности ученых, приводили к конфликтным ситуациям внутри подразделений и между экспертами по анализу и оценке предоставляемых данных.

Для решения перечисленных проблем автором предложена методика и инструмент согласования смысла индикаторов. Описанная ниже методика категоризации знаний о мониторинге, анализе и оценки в сфере науки, в основана на том числе, знаний о результатах категоризации и методе классификации. В качестве главного

[1] В дальнейшем предполагается проведение анализа и оценки результатов и результативности деятельности РАН на четырех институциональных уровнях:

- макроуровень – Российская академия наук в целом;
- мезоуровень – институты Российской академии наук;
- микроуровень – научные коллективы институтов;
- наноуровень – ученые и инженеры.

032-Kozhunova_korr:_Layout 1 18.05.2008 15:24 Page 212

инструмента прояснения смысла используется разрабатываемый семантический словарь и информационные схемы ресурсов системы мониторинга.

Система мониторинга и семантический словарь

В настоящее время, на стадии концептуального проектирования, разработана архитектура словаря, включая ссылки на нормативные, информационные, алгоритмические и другие компоненты системы мониторинга [6]. Архитектура словаря основана на схеме классификации, полученной в результате анализа Приказа № 68 и других нормативных документов и категоризации показателей. Разные категории показателей могут быть связаны между собой родовидовыми и функциональными тезаурусными отношениями (Рис. 1).

При использовании классификационного метода в процессе разработки семантического словаря как средства уточнения смысла индикаторов результатов научной деятельности были рассмотрены и учтены следующие гипотезы и базовые положения:

1. нормативные документы редко содержат явные определения индикаторов и других показателей;
2. иногда значения индикаторов с теми же самыми названиями определены по-разному в различных публикациях;
3. иногда индикаторы научной деятельности неодинаково интерпретируются лицами, принимающими решения, менеджерами, экспертами, специалистами по оценке, лингвистами и IT-специалистами;
4. существуют индикаторы, зависящие от нескольких параметров, изменение которых изменяет их численные значения (например, значения индексов цитирования зависят от глубины ретроспективы используемого массива научных статей). Кроме того, индикаторы могут зависеть от выбора варианта используемого алгоритма вычисления, что может изменить их смысл (например, пункт 2.3 Приказа, приведенный выше – индикатор результативности совместителей имеет несколько вариантов вычисления в силу многозначности приведенного в документе текста);
5. численные значения индикаторов могут зависеть от числа записей в используемых нормативных файлах и содержания этих записей (например, значения индексов цитирования зависят от списка используемых журналов).

Вышеперечисленные аспекты были учтены автором при разработке процедуры уточнения смысла индикаторов, которую предлагается проводить в два этапа. На первом этапе осуществляется встраивание каждого предлагаемого к использованию индикатора в классификационную схему, полученную в результате категоризации показателей. Размещение в схеме позволяет согласовывать предварительное понимание значения индикатора, которое можно извлечь из этой схемы (Рис. 1). В приведенной классификационной схеме предусмотрены отдельные позиции как для индикаторов научной деятельности, так и для технологической и образовательной видов деятельности. Эти виды деятельности приведены в новом Уставе РАН, утвержденным постановлением Правительства РФ от 19 ноября 2007 г. № 785.

На втором этапе происходит уточнение смысла индикаторов посредством использования словарных статей, связанных с нормативными, информационными и алгоритмическими компонентами системы мониторинга.

Основная идея предлагаемой методики получения согласованных индикаторов состоит в разработке итерационной процедуры прояснения их смысла, учитывающей одновременно влияние нормативного, лингвистического и экстралингвистического факторов (например, выбор вариантов вычисления того или иного индикатора, иллюстрации дефиниций индикаторов, ссылки на нормативные документы, точность и ясность дефиниций результатов научной деятельности и т.п.).

Разработанная итеративная процедура согласования смысла индикаторов одновременно использует следующие компоненты системы мониторинга [7, 8]:

- нормативный компонент системы;
- семантический словарь с названиями и определениями видов индикаторов, характеристик, критериев, параметров и экспертных оценок;
- информационный компонент системы;
- библиотеку алгоритмов и программ (алгоритмический компонент системы).

Первый этап согласования смысла индикаторов был протестирован на индикаторах результатов наноуровня. В результате было получено распределение индикаторов результатов для наноуровня в пределах классификационной схемы индикаторов и других категорий показателей, которое будет приведено далее в статье.

032-Kozhunova_korr:_Layout 1 18.05.2008 15:24 Page 213

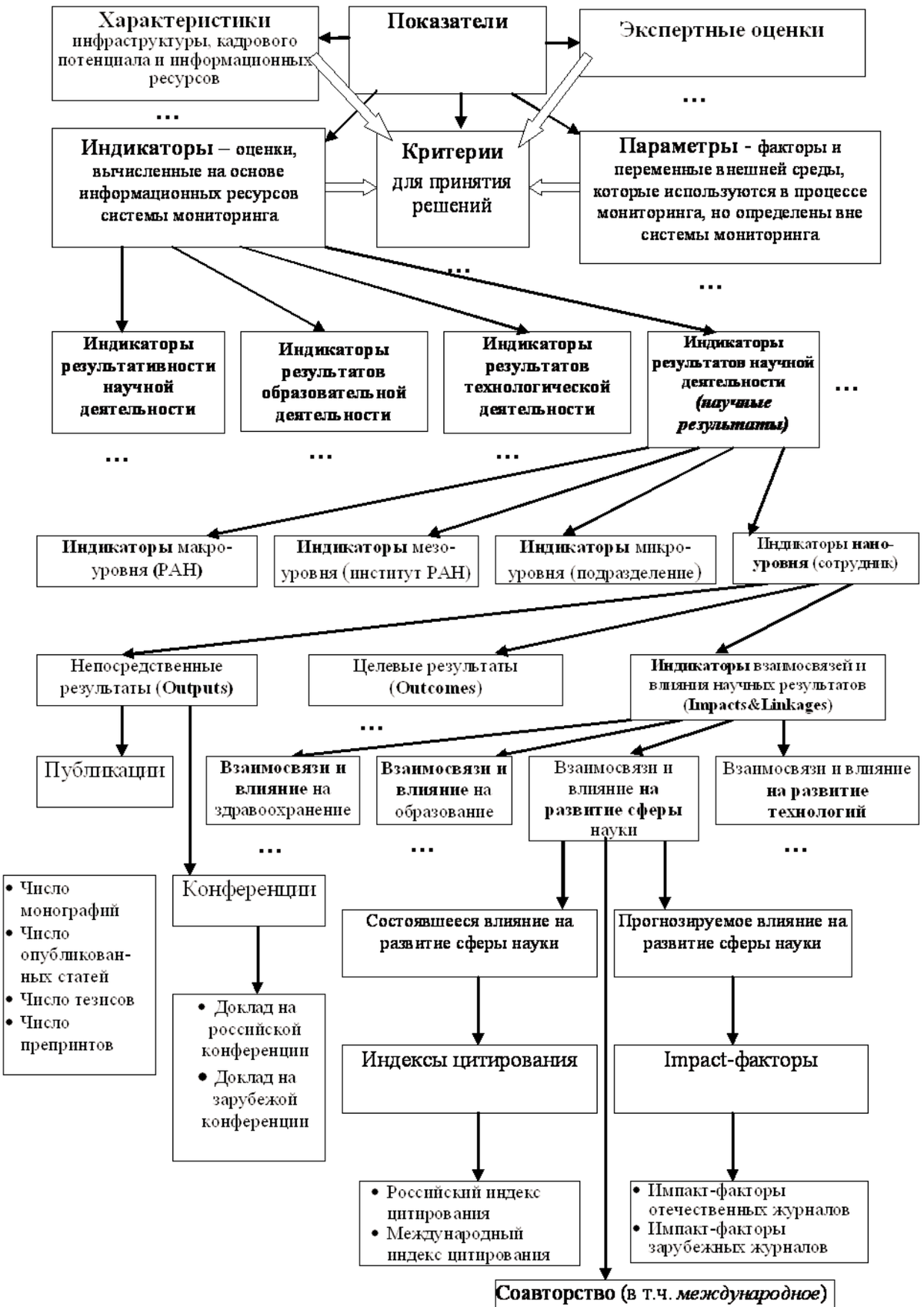


Рис.1. Классификационная схема семантического словаря системы мониторинга: иерархия показателей и отношения между ними

032-Kozhunova_korr:_Layout 1 18.05.2008 15:24 Page 214

По итогам анализа Совместного Приказа № 68 [1] автором был выделен следующий перечень индикаторов результатов для наноуровня (научных сотрудников):

1. Монографии и учебники (единица измерения - число печатных листов);
2. Доклады на российских конференциях (единица измерения – число докладов в течение календарного года);
3. Доклады на международных конференциях (единица измерения – число докладов в течение календарного года);
4. Приглашенные доклады на российских конференциях (единица измерения – число докладов в течение календарного года);
5. Приглашенные доклады на международных конференциях (единица измерения – число докладов в течение календарного года);
6. Разработки новых курсов, выполненные в течение календарного года (единица измерения – суммарное число семестров, вычисленное для каждого нового курса отдельно);
7. Доработки существующих курсов, выполненные в течение календарного года (единица измерения – суммарное число семестров, вычисленное для каждого доработанного курса отдельно);
8. Защиты кандидатских диссертаций в течение календарного года (единица измерения – выписки из решения ВАК или удостоверения кандидатов наук);
9. Защиты дипломов (единицы измерения - выданные дипломы, учитываемые только при условии дальнейшего поступления дипломанта в аспирантуру и/или на работу в научную организацию);
10. Защиты докторских диссертаций в течение календарного года (единица измерения – выписки из решения ВАК или удостоверения докторов наук);
11. Патенты, полученные в течение календарного года (единица измерения – число патентов);
12. Программные продукты, зарегистрированные в течение календарного года (единица измерения – число свидетельств о регистрации программных продуктов);
13. Базы данных, зарегистрированные в течение календарного года (единица измерения – число свидетельств о регистрации баз данных);
14. Топологии микросхем, зарегистрированные в течение календарного года (единица измерения – число свидетельств о регистрации топологий микросхем);
15. Российский индекс цитирования, учитываемый в текущем году (единица измерения – количество ссылок в этом индексе; измеряется на основе импакт-факторов журналов, в которых опубликованы статьи);
16. Международный индекс цитирования, учитываемый в текущем году (единица измерения – количество ссылок в этом индексе; измеряется на основе импакт-факторов журналов, в которых опубликованы статьи);
17. Статьи в российских или зарубежных журналах с impact-фактором $\geq 0{,}2$ (единица измерения – число статей в календарном году);
18. Статьи в зарубежных журналах с impact-фактором $\leq 0{,}2$ или без impact-фактора (единица измерения – число статей в календарном году);
19. Статьи в российских журналах, включенные в Перечень ВАК для кандидатских диссертаций (единица измерения – число статей в календарном году);
20. Статьи в российских журналах, включенные в Перечень ВАК для докторских диссертаций (единица измерения – число статей в календарном году).

В пределах классификационной схемы, полученной в результате категоризации, каждый из двадцати перечисленных индикаторов отнесен к одной из категорий, указанных на схеме (Рис.1). В результате проведенной классификации этих двадцати индикаторов автором было получено следующее распределение (для каждой из ниже перечисленных категорий указан полный или сокращенный путь к этой категории, начиная с вершины схемы на рис.1, и приведен список индикаторов, относящихся к этой категории):

1. Показатели → Индикаторы → Индикаторы результатов научной деятельности → Индикаторы наноуровня → Непосредственные результаты → Публикации:

- Монографии и учебники;
- Статьи в российских журналах, включенные в Перечень ВАК для кандидатских диссертаций;
- Статьи в российских журналах, включенные в Перечень ВАК для докторских диссертаций.

2. Показатели → ...→ Конференции:

- Доклады на российских конференциях;
- Доклады на международных конференциях;
- Приглашенные доклады на российских конференциях;
- Приглашенные доклады на международных конференциях.

3. Показатели → Индикаторы → Индикаторы результатов научной деятельности → Индикаторы наноуровня → ...→ Индексы цитирования:

032-Kozhunova_korr:_Layout 1 18.05.2008 15:24 Page 215

• Российский индекс цитирования, учитываемый в текущем году;
• Международный индекс цитирования, учитываемый в текущем году.

4. Показатели → Индикаторы → Индикаторы результатов научной деятельности → Индикаторы наноуровня → ...→ Impact-факторы:

• Статьи в российских или зарубежных журналах с impact-фактором ≥ 0,2;
• Статьи в зарубежных журналах с impact-фактором ≤ 0,2 или без impact-фактора.

5. Показатели → ...→ Индикаторы результатов образовательной деятельности:

• Разработки новых курсов, выполненные в течение календарного года;
• Доработки существующих курсов, выполненные в течение календарного года;
• Защиты кандидатских диссертаций в течение календарного года;
• Защиты дипломов;
• Защиты докторских диссертаций в течение календарного года.

6. Показатели → ...→ Индикаторы результатов технологической деятельности:

• Патенты, полученные в течение календарного года;
• Программные продукты, зарегистрированные в течение календарного года;
• Базы данных, зарегистрированные в течение календарного года;
• Топологии микросхем, зарегистрированные в течение календарного года.

Распределение перечисленных индикаторов по шести категориям (рубрикам) классификационной схемы наглядно показывает, что все согласованные и утвержденные индикаторы наноуровня для результатов научной деятельности относятся к категории «Непосредственные результаты» или к категории «Индикаторов взаимосвязи и влияния научных результатов на сферу науки». При этом полностью отсутствуют индикаторы таких значимых категорий как «Целевые результаты» и «Индикаторы результативности научной деятельности». Это говорит, прежде всего, о неоднородности утвержденного Приказом № 68 списка индикаторов результатов научной деятельности, а также о том, что в этом нормативном документе были проигнорированы достаточно весомые при формировании согласованных оценок категории. На практике, при проведении эксперимента, кратко описанного в статье, это привело к затруднению принятия решений экспертами при вынесении предварительных и итоговых оценок результативности научной деятельности сотрудников Российской Академии наук, а также к неоднозначным трактовкам индикаторов и их значений. Следствием этого были конфликтные ситуации между экспертами и оцениваемыми сотрудниками и не всегда адекватный учет значимых научных результатов (например, у таких категорий сотрудников как совместители, молодые специалисты, руководители подразделений и т.п.)[2].

Заключение

В статье представлено краткое описание эксперимента по оценке индивидуальных показателей результативности научной деятельности, проведенного в Российской Академии наук в первом полугодии 2007 года, а также проблемы, выявленные по его итогам.

В результате анализа соответствующих нормативных документов, являющихся основанием для проведения эксперимента, и итогов самой процедуры этого мероприятия была выявлена необходимость в разработки методики согласования смысла индикаторов результатов научной деятельности (базовых терминов, одобренных нормативными документами и использованных в ходе эксперимента). В качестве основного метода автором был предложен классификационный метод, основанный на итеративной процедуре уточнения смысла индикаторов с последовательной категоризацией этих терминов и поиском их дефиниций и иллюстраций значений при помощи компонент системы мониторинга. Кроме того, была выявлена необходимость в разработки инструмента, обеспечивающего проведение мониторинга, анализа и оценку результативности научной деятельности в Российской академии наук. Это семантический словарь, проектируемый и разрабатываемый в настоящее время параллельно с созданием системы мониторинга в ИПИ РАН и ЦЭМИ РАН. Классификационная схема словаря, встроенная в него для проведения первого этапа классификации, позволила выявить ряд недочетов в нормативных документах, содержащих перечень индикаторов для проведения оценки результативности научных сотрудников, и наметить пути решения проблемы согласования понимания смысла индикаторов.

[2] На основе данных эксперимента по оценке результативности индивидуальных показателей результативности научной деятельности в ИПИ РАН.

032-Kozhunova_korr:_Layout 1 18.05.2008 15:24 Page 216

Список литературы

1. Совместный приказ Министерства образования и науки РФ, Министерства здравоохранения и социального развития РФ и Российской академии наук № 273/745/68 от 3 ноября 2006г. - http://www.ras.ru/news/shownews.aspx?id= 328a9bb8-b7ba-4275-93f6-f63698c471d9&_Language=en.

2. Постановление Правительства РФ от 22.04.2006 года № 236 «О реализации в 2006-2008 годах пилотного проекта совершенствования системы оплаты труда научных работников и руководителей научных учреждений и научных работников научных центров РАН» (http://www.government.ru/data/news_print.html?he_id=103&news_id=21148).

3. Analytical Perspectives. Budget of the United States Government, Fiscal Year 2007. - Washington, DC: U.S. Government printing office, 2006.

4. Минин В.А. Мониторинг научных исследований российских ученых. В кн.: Российский фонд фундаментальных исследований: десять лет служения российской науке.- М.: Научный мир, 2003.- С. 295-314.

5. National Science Board, Science and Engineering Indicators – 2006. Two volumes. Arlington, VA: National Science Foundation, 2006.

6. Кожунова О.С., Зацман И.М. Прагматические аспекты создания семантического словаря терминов информационного мониторинга // Труды международной конференции Диалог-2007 «Компьютерная лингвистика и интеллектуальные технологии».- М.: Изд-во РГГУ, 2007. - С. 278-285.

6. Зацман И.М. Информационные ресурсы для систем мониторинга в сфере науки // Системы и средства информатики. Вып. 15.- М.: Наука, 2005 (б).- С. 288-318.

7. Зацман И.М. Полидоменные модели в системах оценки инновационного потенциала и результативности научных исследований // Тр. конференции Диалог-2006 «Компьютерная лингвистика и интеллектуальные технологии». - М.: РГГУ, 2006. - С. 178-183.

033-Kozlova:_Layout 1 18.05.2008 15:25 Page 217

‘ЗАСТАВИТЬ’ ИЛИ ‘РАЗРЕШИТЬ’: АНАЛИЗ СЕМАНТИКИ КАУЗАТИВНЫХ ГЛАГОЛОВ

‘CAUSE’ OR ‘ENABLE’: ANALYSIS OF CAUSATIVE VERBS SEMANTICS

Козлова А.В. *(avkozlova@rambler.ru),* ***Люткова Е.А.*** *(katjal@philol.msu.ru),*
Федорова О.В. *(olga.fedorova@msu.ru)*
Московский государственный университет имени М.В. Ломоносова

В работе приводятся данные экспериментального исследования семантики русских каузативных глаголов в рамках Модели динамики сил. В зависимости от соотношения параметров каузативной ситуации (стремление каузируемого к достижению конечного состояния, оппозиция каузатора и каузируемого и наличие конечного состояния) выделяются три типа каузативных отношений: CAUSE, ENABLE и PREVENT.

1. Введение

Под каузативными глаголами (от лат. *causa* ‘причина’) в настоящей работе понимаются глаголы типа ‘вынуждать’, ‘разрешать’, ‘запрещать’ и т.п. (синтаксические каузативы по [Kulikov 2001]), обозначающие в языке каузативную ситуацию – макроситуацию, состоящую из “по меньшей мере двух микроситуаций, связанных между собой отношением каузации или причинения” [Недялков, Сильницкий 1969]. Отношение каузации имеет место, если один из фактов действительности изменяется в результате воздействия на него другого.

(1) *Полицейский вынудил девушку подойти к мужчине.*

2. Различные подходы к анализу семантики каузативных глаголов

Каузация считается многими исследователями концептуальным примитивом (см. [Wolff 2007]). Такое представление о каузации имеет своих сторонников как среди приверженцев функционально-типологического подхода к анализу лингвистических феноменов, так и среди последователей формальной традиции в описании языка. С точки зрения функционализма каузативное отношение является универсальным элементарным понятием, имеющим определенное выражение в конкретном языке [Shibatani, Pardeshi 2002]. С точки зрения формальной семантики, каузативное отношение предстает как логическое отношение между событиями или подсобытиями, интерпретируемое в терминах временных интервалов или импликативных следствий [Lombard 1985], [Hale, Keyser 1993].

Эта точка зрения представляется спорной. Так, многие исследователи выявляют различные семантические типы каузации, находящие регулярное отражение в грамматике естественного языка (см., например, [Kulikov 2001], [Ginet 1990], [Kratzer 2005]).

В работе [Kulikov 2001] обобщаются исследования по типологии каузативов и различаются следующие виды каузативных глаголов в зависимости от типа каузации:

• прямой vs. непрямой каузатив (критерий: физическое воздействие каузатора на каузируемого), ср. (2) из зырянского языка:

(2)

puk-	*puk-t-*	*puk-öd-*	*[Лыткин 1957: 105]*
сидеть	садить	вынудить сесть	

• пермиссив vs. фактитив (критерий: разрешение каузатора совершить действие, отсутствие собственно каузации со стороны каузатора). Формальное выражение пермиссива и фактитива во многих языках совпадает [Kulikov, Nedjalkov 1992]. Ср. (англ.) (3):

(3) *let sleep ~ not to cause not to be awake*

В работе [Недялков, Сильницкий 1969] утверждается, что семантически все каузативные связки (включая каузативные глаголы) могут, наряду с собственно каузацией, выражать каузирующее состояние ((4): zornrot – инструментальная каузативная связка) и/или каузируемое состояние (приказ, разрешение – результативные каузативные связки).

033-Kozlova:_Layout 1 18.05.2008 15:25 Page 218

(4) *Er* *war* *zornrot.* [Недялков, Сильницкий 1969: 9]
Он был красным от гнева

	Нерезультативные	Результативные
Неинструментальные	Одноконстантные заставить (K)	Двухконстантные испугать (K Sj)
Инструментальные	Двухконстантные приказать (Si K)	Трехконстантные подозвать (Si K Sj)

Таблица 1. Семантическая классификация каузативных глаголов русского языка по [Недялков, Сильницкий 1969] (каузирующее состояние – Si, отношение каузации – K, каузируемое состояние – Sj)

Однако, как замечают сами авторы этой классификации, она не учитывает, что кроме констант Si, Sj и K в смысловом содержании глаголов каждой группы есть и другие семантические признаки, различающие их.

3. Модель динамики сил как универсальный способ классификации каузативных глаголов

В рамках когнитивной лингвистики Л. Талми [Talmy 1988] было предложено выделить Модель динамики сил (МДС), которая описывает взаимодействие между сущностями при помощи понятия силы и является значимой для формирования значения ситуации. Согласно этой теории, концепт каузации не является семантическим примитивом. Каждая каузативная ситуация представляет собой противостояние двух неравных сил (каузатора и каузируемого). Направление этих сил и соотношение между ними определяют характер каузативной ситуации.

Таким образом, предполагается, что высказывания на естественном языке либо маркированы, либо нейтральны с точки зрения категории динамики сил:

(5) а. *Дверь закрыта.* [Talmy 1988: 62]
б. *Дверь не может открыться.*

Предложение (5а) является нейтральным в сило-динамическом аспекте, так как в ситуации, описываемой этим предложением, нет противопоставления сил. Предложение (5б) выражает некоторое значение категории динамики сил. Очевидно, что субъект (*дверь*) имеет тенденцию к тому, чтобы совершить действие (*открыться*), однако существует сила, препятствующая совершению этого действия (например, дверь может быть заперта или дверь могла захлопнуться).

Две выделенные силы кодируются языком в зависимости от ситуации. Сила, которая находится в фокусе, называется в работе [Talmy 1988] протагонистом (Agonist); сила, которая противостоит ей, – антагонистом (Antagonist). В (5б) протагонистом является *дверь*, антагонистом – сила, которая препятствует тому, чтобы дверь открылась. Третий релевантный фактор – соотношение между силами. По определению, если ситуация маркирована с точки зрения теории динамики сил, то силы не равны. Таким образом, если силы будут одинаково сильными, то ситуация не рассматривается с точки зрения динамики сил. В (5б) антагонист сильнее, так как он держит дверь закрытой. Результат противодействия сил зависит от внутренней тенденции и соотношения между силами. Результатом может быть: действие или отсутствие действия. В ситуации (5б) дверь остается закрытой (отсутствие действия). Делается вывод, что ситуации, в которых протагонист сильнее, выражаются конструкциями ‘X случилось, несмотря на Y’; ситуации, в которых сильнее антагонист – ‘X случилось из-за Y’.

В дальнейшем теория динамики сил разрабатывалась некоторыми другими исследователями. Так, в работе [Wolff, Ventura 2003] на основе модели динамики сил выделяются три типа каузативных отношений: CAUSE, ENABLE и PREVENT[1], каждый из которых характеризуется определенным набором значений следующих трех параметров: стремление каузируемого к достижению конечного состояния, оппозиция каузатора и каузируемого и наличие конечного состояния[2] (см. Таблицу 2). Под стремлением (tendency) понимается “предрасположенность пациенса к конечному результату, обусловленная внутренними особенностями или деятельностью пациенса” [Wolff, Ventura 2003: 824]. Оппозиция имеет место, когда “сила аффектора, влияющая на пациенса, противоречит стремлению пациенса” [Wolff, Ventura 2003: 824].

[1] Перевод этих терминов на русский язык может изменить оттенки смысла, вкладываемого авторами рассматриваемой статьи в понятия, стоящие за этими терминами. О соотношениях значений русских и английских каузативных глаголов см. [Wolff, Ventura 2003].

[2] Выделение только этого параметра, хотя и самого очевидного с прагматической точки зрения, было бы недостаточным для классификации. Каузативные глаголы являются глаголами с пропозициональным актантом. С точки зрения содержащихся в значении глаголов импликаций относительно истинности подчиненной предикации (ПП) каузативные глаголы русского языка делятся на несколько типов:

033-Kozlova:_Layout 1 18.05.2008 15:25 Page 219

	Стремление каузатора к достижению конечного состояния	Оппозиция каузатора и каузируемого	Наличие конечного состояния
CAUSE	Нет	Есть	Есть
ENABLE	Есть	Нет	Есть
PREVENT	Есть	Есть	Нет

Таблица 2. Представление отношений CAUSE, ENABLE, PREVENT с помощью МДС

Психологическую реальность рассматриваемого когнитивного подхода к представлению отношения каузации в сознании носителя языка можно проверить при помощи психолингвистических экспериментов. В экспериментах Ф. Вольфа и соавторов испытуемым предлагаются визуальные стимулы (анимации, смоделированные согласно соответствующему набору значений параметров), которые необходимо описать конструкцией с синтаксическим каузативом. Результаты экспериментов на материале английского языка показали, что носители языка верно (т.е. используя ожидаемый каузативный глагол) определяют тип каузативной ситуации, смоделированной согласно соответствующему набору значений параметров. Таким образом, подтвердилось, что ментальная репрезентация концепта каузации может объясняться МДС (см. Таблицу 3).

		Оценка испытуемого (используемый глагол)					
		Cause		Help		Prevent	
		Кол-во	%	Кол-во	%	Кол-во	%
Тип ситуации	CAUSE	**17**	**94% (.236)**	1	6%	0	0%
	ENABLE	2	11%	**16**	**89% (.323)**	0	0%
	PREVENT	0	0%	0	0%	**18**	**100% (0)**

Таблица 3. Описание каузативной ситуации носителями английского языка ([Wolff 2007])

4. Эксперимент

Целью настоящей работы было выявить релевантность МДС при описании отношения каузации в когнитивной системе носителей русского языка. А именно, выяснить, насколько выделение в семантике русских каузативных глаголов трех описанных выше параметров и постулирование, в соответствии со значениями этих трех параметров, существования в русском языке трех типов каузативных ситуаций (CAUSE, ENABLE и PREVENT) соответствует реальному устройству каузативных отношений в сознании носителей русского языка.

В соответствии с поставленными задачами был разработан и проведен эксперимент на материале русского языка. Ранее исследований семантики русских каузативных глаголов в рамках МДС не проводилось. (В работе [Wolff, Ventura 2003] описано исследование, в котором участвовали носители русского языка, однако его цели отличались от целей нашего эксперимента.)

4.1. Стимульный материал

В качестве методики проведения эксперимента был выбран опросник. Это опосредованный вид анализа: такая методика позволяет изучать и моделировать процессы анализа предложения при восприятии уже на основе результатов этого анализа. В эксперименте использовались стимулы в визуальной модальности – анимации. Стимульный материал такого вида максимально отвечает целям эксперимента: таким образом испытуемому можно предъявить ситуацию, сгенерированную в соответствии с заданным типом каузации.

В эксперименте контролировался тип каузативной ситуации согласно МДС. Были взяты 3 сюжета (‘вентиляторы и лодка’, ‘полицейский и девушка’, ‘вентиляторы и девушка на круге’), каждый из которых был представлен в трех вариантах:

1. тип CAUSE (у пациенса нет стремления к достижению результата, есть оппозиция аффектора и пациенса, результат достигается);

(6) а. ‘заставить’ <И, Л>
б. ‘помочь’ <И, 0>
в. ‘помешать’ <Л, И>
г. ‘дать возможность’ <0, Л>
д. ‘разрешить’ <0, Л>
[Зализняк 1988]

033-Kozlova:_Layout 1 18.05.2008 15:25 Page 220

2. тип ENABLE (у пациенса есть стремление к достижению результата, нет оппозиции аффектора и пациенса, результат достигается);

3. тип PREVENT (у пациенса есть стремление к достижению результата, есть оппозиция аффектора и пациенса, результат не достигается).

№ блока	Тип	Анимация
1	CAUSE	Вентиляторы заставили лодку доплыть до полосатого конуса.
	ENABLE	Вентиляторы помогли лодке доплыть до полосатого конуса.
	PREVENT	Вентиляторы помешали лодке доплыть до полосатого конуса.
2	CAUSE	Полицейский вынудил девушку подойти к мужчине.
	ENABLE	Полицейский разрешил девушке подойти к мужчине.
	PREVENT	Полицейский запретил девушке подходить к мужчине.
3	CAUSE	Вентиляторы заставили девушку доплыть до конуса.
	ENABLE	Вентиляторы помогли девушке доплыть до конуса.
	PREVENT	Вентиляторы помешали девушке доплыть до конуса.

Таблица 4. Список экспериментальных блоков

Материалом эксперимента стали 9 экспериментальных анимаций, которые составили 3 экспериментальных блока, и 3 отвлекающих анимации.

В ходе эксперимента фиксировался выбор каузативного глагола из группы CAUSE, из группы ENABLE или из группы PREVENT (оценка сгенерированной каузативной ситуации).

4.2. Процедура проведения эксперимента

В эксперименте приняло участие 60 человек в возрасте от 18 до 50 лет. Испытуемым предлагалось просматривать анимации, в которых были показаны ситуации, смоделированные в соответствии с той или иной комбинацией значений трех выделенных параметров, и затем выбирать предложение с тем каузативным глаголом, который, по мнению испытуемого, лучше описывает ситуацию.

Испытуемые тестировались индивидуально, каждый эксперимент продолжался примерно 10 минут. Каждому участнику предлагался один из трех экспериментальных листов. На экране компьютера запускалась анимация, испытуемый просматривал ее, и затем, на экране появлялись три предложения, из которых испытуемый выбирал то, которое, на его взгляд, максимально точно описывало просмотренную анимацию. Порядок предложений был сбалансирован.

(6) *Полицейский вынудил девушку подойти к мужчине.*
Полицейский разрешил девушке подойти к мужчине.
Полицейский запретил девушке подходить к мужчине.

Испытуемый выбирал один из вариантов, произнося его вслух, и, нажимая на соответствующую клавишу, переходил к следующей анимации.

4.3. Результаты и направление дальнейших исследований

• Проведенное исследование показало, что МДС недостаточно четко описывает концепт каузации в когнитивной системе носителей русского языка. Ситуация, смоделированная согласно соответствующему набору значений параметров, не всегда определяется глаголом предполагаемого типа (см. Таблицу 5);

• проведенный статистический анализ с использованием критерия Фридмана показал значимое различие в определении трех ситуаций (χ^2*эмп* = 24,175 для $P \leq 0,05$);

• наиболее четко определяется по заданным параметрам ситуация PREVENT. Имеется значимое статистическое различие в определении этой ситуации по сравнению с двумя другими (CAUSE vs. PREVENT (χ^2*эмп* = 26,82 для $P \leq 0,05$), ENABLE vs. PREVENT (χ^2*эмп* = 16,5 для $P \leq 0,05$)). На наш взгляд, это связано с наличием в ситуации PREVENT фактора отсутствия результата;

033-Kozlova:_Layout 1 18.05.2008 15:25 Page 221

• при описании типа каузативной ситуации в русском языке представляется необходимым ввести фактор одушевленности / неодушевленности каузатора и каузируемого для объяснения сильного смешения ситуаций CAUSE и ENABLE (в определении ситуаций CAUSE и ENABLE отсутствует значимое различие (χ^2*эмп* = 1,65 для P ≤ 0,05). По результатам нашего эксперимента, если каузатор неодушевленный, то выбирается глагол из группы CAUSE; если каузируемый одушевленный, то выбирается глагол из группы ENABLE. Это заключение, однако, нуждается в дополнительной проверке.

		Оценка испытуемого (используемый глагол)					
		Заставили / вынудил		Помогли / разрешил		Помешали / запретил	
		Кол-во	%	Кол-во	%	Кол-во	%
Тип ситуации	CAUSE	**29**	**48,33%**	29	48,33%	3	3,33%
	ENABLE	19	32%	**36**	**60,00%**	4	8%
	PREVENT	5	8%	0	0%	**55**	**92%**

Таблица 5. Описание каузативной ситуации носителями русского языка

Список литературы

1. Зализняк Анна А. О понятии импликативного типа // Логический анализ языка. М.: Знание и мнение, 1988. С. 107-121.

2. Лыткин В.И. Принудительный залог в пермских языках // Ученые записки Удмуртского НИИ. №18. 1957. P. 93-113.

3. Недялков В.П., Сильницкий Г.Г. Типология каузативных конструкций // Типология каузативных конструкций. Морфологический каузатив. Ленинград: “Наука”,1969. С. 5-19.

4. Ginet C. On action // Cambridge: Cambridge University Press, 1990.

5. Kratzer A. Building resultatives // Event arguments in Syntax, Semantics, and Discourse. Tübingen: Niemeyer, 2005. P. 177-212.

6. Kulikov L. Causatives // Language typology and language universals. An international handbook. Berlin: 2001. Vol. 2. P. 886-898.

7. Kulikov L.I., Nedjalkov V.P. Questionnaire zur Kausativierung // Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung. 45.2 Berlin: Akademie-Verlag, 1992. P. 137-149.

8. Lombard B. How not to flip the prowler: Transitive verbs of action and the identity of actions // Actions and Events: Perspectives on the Philosophy of Donald Davidson. Oxford: Basil Blackwell, 1985. P. 268-281.

9. Shibatani M., Pardeshi P. The causative continuum // The Grammar of Causation and Interpersonal Manipulation. Amsterdam: John Benjamins, 2002. P. 85-126.

10. Talmy L. Force dynamics in language and cognition // Cognitive Science, 1988. № 12. P. 49-100.

11. Wolff P. Representing causation // Journal of Experimental Psychology: General, 2007. № 136. P. 82-111.

12. Wolff P., Ventura T. When Russians learn English: How the meaning of causal verbs may change // Proceedings of the Twenty-Seventh Annual Boston University Conference on Language Development. Boston: Cascadilla Press, 2003. P. 822-833.

034-Kolmogorova:_Layout 1 18.05.2008 15:47 Page 222

ЭВОЛЮЦИЯ ФОРМ РЕЧЕВОГО АРГУМЕНТАТИВНОГО ПОВЕДЕНИЯ КАК ОДИН ИЗ АСПЕКТОВ СТАНОВЛЕНИЯ КОММУНИКАТИВНОЙ КОМПЕТЕНЦИИ ЯЗЫКОВОЙ ЛИЧНОСТИ

EVOLUTION OF ARGUMENTATIVE LANGUAGE BEHAVIOUR PATTERNS AS AN ASPECT OF COMMUNICATIVE COMPETENCE GENESIS

Колмогорова А.В. *(nastiakol@mail.ru)*
Кузбасская государственная педагогическая академия

В статье обобщаются результаты экспериментальной работы по выявлению форм речевого аргументативного поведения представителей русского лингвокультурного сообщества, принадлежащих трём возрастным группам. Делается вывод о том, что аргументация есть речевой жанр, формы реализации которого осваиваются языковыми личностями в онтогенезе.

0. Введение
Понятия речевой деятельности и речевого поведения

Для лингвистики последнего десятилетия характерна смена, так называемого, «протофеномена» (Холодная 1997) − того ключевого явления (факта, экспериментальной ситуации и т.д.), в котором манифестируется та или иная теория, и которое, в свою очередь, задает некоторые исходные теоретические ориентации в изучении природы того или иного аспекта реальности. На смену протофеномену «язык как система» приходит протофеномен «речь», следовательно, лингвистика языка сменяется лингвистикой речи, объектом которой выступает совокупность индивидуальных, частных и мгновенных проявлений речевой деятельности (Соссюр 1998). Термин «речевая деятельность» претерпел достаточное количество интерпретаций: классическое определение «язык +речь» (Соссюр 1998); сложное или простое волевое действие, направленное на формулирование мысли в слове (Выготский 2007); процессы говорения и понимания, обусловленные сложной психофизиологической речевой организацией индивида (Щерба 1974); один из аспектов более общего понятия деятельности, включающего в себя такие обязательные составляющие, как мотив, цель, потребность, действия и средства их осуществления (Леонтьев 1997).

Однако термин «речевая деятельность», несмотря на некоторую диффузность содержательного наполнения, предусматривает определённый ракурс рассмотрения процесса актуализации языковой способности человека: фокусируя внимание на исключительных качествах человека как существа, снабжённого определёнными априорными категориями, способностями (например, языковой способностью), отделяющими его от других живых существ (Декарт, Кант), сочетание «речевая деятельность» предполагает наличие только одного «центра притяжения», одного активного и доминирующего субъекта – говорящего/слушающего, действующего при помощи языка, руководствуясь своими собственными целями и мотивами. Однако, как показывает развитие современных гуманитарных и естественных наук, нельзя забывать о том, что человек суть биологический организм, существование которого тесно связано с окружающей био-социо-культурной средой, к которой человек, как и любой другой организм, должен адаптироваться, гибко реагируя на те её изменения, которые существенны для его жизнедеятельности. Развитие в 20 веке биосемиотики, теории самоорганизующихся систем, теории автопойезиса способствует смещению исследовательского интереса с процессов действия человека в мире на процессы взаимодействия человека с этим миром, где языковая способность приобретает статус важнейшего адаптивного механизма, поскольку позволяет homo loquens выйти за пределы своего собственного когнитивного опыта, образа своей субъективной реальности, для того, чтобы моделировать общий мир с другими членами сообщества, мир, в котором можно сосуществовать и иметь взаимный доступ (Матурана 2002:7). В этой связи в научный дискурс был введён термин «речевое поведение», предусматривающий рассмотрение речевой активности человека через призму двух ключевых понятий: взаимодействие и выбор.

Л.С.Выготский определял поведение как процесс взаимодействия между организмом и средой (Выготский 2007), детерминируемый характером реагирования организма на сигналы извне. Под речевым поведением понимается стереотипичное и индивидуальное проявление заавтоматизированного, неосознаваемого,

034-Kolmogorova:_Layout 1 18.05.2008 15:47 Page 223

интуитивного выбора отправителем текста речевых сигналов актуализации скрытых нюансов смысла (Матвеева 1992), словесно выраженная часть эмпирически наблюдаемого и воспринимаемого адресатом внешнего проявления коммуникативной деятельности, последовательность речевых поступков (Борисова 2003:67). Мы определяем речевое поведение как процесс взаимодействия человека с окружающей био-социо-культурной средой, осуществляемый посредством интуитивного, основанного на предыдущем опыте, выбора субъектом тех или иных форм реализации языковой способности.

Одной из форм речевого поведения является аргументативное поведение. Уточним, что, будучи рассмотрена достаточно широко, аргументация является имманентным свойством всякого общения, понимаемого как ***глубинный процесс со-бытия двух и более людей в процессе моделирования, создания ими общего мира на основе пред-общности речежизненного опыта*** (потенциальная общность прецедентных феноменов, культурных стереотипов (Красных 2003), близость био-социо-культурных структур знания, определяемых нами как значения языковых знаков (Колмогорова 2006)).

Однако, несомненно, есть определённый набор форм реализации языковой способности, типично используемый представителями того или иного национально-лингво-культурного сообщества в той или иной ситуации взаимодействия. В современной лингвистической литературе отмечается вновь возросший интерес к обсуждению уже поднимавшейся в начале 20 в. проблемы жанров речи, как типических форм осуществления речевого поведения, своеобразных фреймов, сценариев. М.М.Бахтин подчёркивал, что лингвистика должна, прежде всего, изучать формы и типы речевого взаимодействия в связи с конкретными условиями его, а также формы отдельных высказываний (жанры речевых выступлений), и, уже исходя из полученных данных, должна пересмотреть формы языка в их обычной лингвистической трактовке (Бахтин 1998). К.Ф.Седов рассматривает жанровое мышление как основу и принцип дискурсивного мышления вообще, при этом эволюция жанрового мышления во многом предопределяет эволюцию языковой личности и её коммуникативной компетенции в целом (Седов 2003:237).

Мы считаем возможным рассмотреть речевое аргументативное поведение (далее – РАП) как один из возможных речевых жанров, имеющий, с одной стороны, некий национально и культурно обусловленный репертуар речевых форм, используемый представителями того или иного лингвокультурного сообщества в типизированной ситуации взаимодействия, а, с другой стороны, - определённую динамику и закономерности становления, эволюционирования в онтогенезе.

1. Экспериментальная работа

1.1. Условия эксперимента

В целях выявления специфических форм реализации РАП в русском лингвокультурном сообществе, а также выявления динамики и закономерностей эволюционирования данных форм мы использовали такой вид исследовательской работы, как эксперимент.

Трём группам испытуемых (1 группа - школьницы 14-15 лет (42 чел.), 2 группа - студентки 16-17 лет (51 чел.), 3 группа - студентки 20-24 года (43 чел)) был предъявлен небольшой текст, следующего содержания:

«Игорь рос тихим, спокойным парнем. Родители умерли рано, а его воспитанием занималась старшая сестра Варя. Варя делала всё, чтобы мальчик не чувствовал себя сиротой: покупала хорошие вещи и игрушки, водила в цирк и в кино, старалась окружить мальчика теплотой и заботой. Конечно, для этого сестре приходилось много работать, на личную жизнь времени не хватало, да и к чему это: все мысли были о младшем брате. Когда Игорь поступил в институт, Варя случайно узнала о его романе с женщиной, которая была на 10 лет старше его, у неё уже был ребёнок от первого брака. Как ни пыталась Варя убедить брата не связывать свою жизнь с этой женщиной, брат настоял на своём: Игорь и Лена (так звали избранницу) поженились. Семейная жизнь не заладилась с самого начала. Игорь подозревал жену в изменах, дело иногда доходило до драки. Видя всё это, Варя тяжело заболела – перестали слушаться ноги. Вскоре Лена исчезла в неизвестном направлении, оставив своего ребёнка от первого брака Игорю. Молодой человек сначала запил, а затем, одумавшись, решил целиком посвятить себя карьере. Чтобы помочь брату, воспитанием чужого ребёнка занялась не переставшая страдать от болей в суставах Варя».

Каждой группе испытуемых было предложено одно и то же задание: «Выскажитесь, пожалуйста, в поддержку тезиса № 1 «Варя – хорошая» или в поддержку тезиса № 2 «Варя – плохая». Обоснуйте свою точку зрения». На выполнение задания в письменной форме испытуемые в среднем тратили не более 20-15 минут.

Подчеркнём, что мы намеренно подобрали речевой отрывок, затрагивающий сферу повседневного существования, межличностных взаимоотношений, а не, скажем, профессиональных, или других, более узких, сфер, чтобы вовлечь в поле анализа область так называемой жизненной идеологии (по М.М.Бахтину), составляющей основу коллективного когнитивного пространства коммуникантов (Красных 2003).

034-Kolmogorova:_Layout 1 18.05.2008 15:47 Page 224

1.2. Результаты эксперимента
1.2.1. Типология форм речевого аргументативного поведения

Анализ примеров РАП позволил выявить определённый репертуар его форм. Выявленные формы или типичные способы реализации языковой способности в коммуникативной ситуации аргументации можно разделить на 3 группы:

1) формы РАП, в которых доминирует опора на исходную ситуацию, опорный текст;

2) формы РАП, в которых доминирует опора на моральные, нравственные ценности национально-лингвокультурного сообщества;

3) формы РАП, которые отражают индивидуальные особенности речевой личности аргументатора: а) формы РАП, обусловленные индивидуальными психологическими особенностями речевой личности аргументатора; б) формы РАП, обусловленные определённым уровнем владения аргументатором средствами выражения используемого языка.

Рассмотрим каждую из групп подробнее.

Формы РАП, в которых доминирует опора на исходную ситуацию, опорный текст, включают в себя следующие разновидности:

1) подробное воспроизведение ситуации близко к тексту;

2) краткое воспроизведение ситуации с использованием исходной лексики (*не оставила брата, сделала всё, чтобы он не чувствовал себя сиротой, а когда его бросила жена, она помогла брату, занимаясь воспитанием чужого ребёнка)*;

3) резюмирование ситуации с расстановкой акцентов, осуществляемой при помощи: а) введения в речь аргументативных маркеров, таких, как *ещё, уже, даже; хотя; безусловно; ведь, всё-таки; просто, в конце концов; всё же* и т.д. (***Но*** *когда брат её женился, от первого срыва у неё стали проблемы с ногами.* ***И тут же*** *она* ***уже*** *больная,* ***опять*** *берёт воспитывать ребёнка,* ***только уже*** *не своего,* ***а*** *совсем чужого,* ***и даже*** *ей не родного);* б) при помощи замены лексических единиц с менее выраженным языковым и речевым аргументативным потенциалом (Колмогорова 2006; 2007) на языковые единицы, в которых такой потенциал более выражен: *она пожертвовала собой, своей личной жизнью, интересами, увлечениями* - более нейтральное *делала всё* (из исходного текста) заменяется сочетанием *пожертвовала собой*, которое, с точки зрения языкового аргументативного потенциала, актуализирует в русском языковом сознании представление о прототипической, нормативной модели поведения человека, а, будучи употреблено в речи, актуализирует скрытый социальный императив «и такой человек достоин любви и уважения»;

4) широкое обобщение ситуации (*всё не очень хорошо; она всегда помогала брату и всё такое*);

5) воспроизведение ситуации с попыткой предположения внутренних причин поведения героев, сопровождаемое конструкциями «кажется», «я не знаю», «возможно», «мне представляется», « думаю»: *Наверное, по её мнению, этот брак был не очень разумным; кажется, что у неё не было сильного уважения к себе;*

6) градуирование ситуации (*она скорее хорошая, чем плохая; её можно назвать не только хорошей, но ещё и смелой, отважной женщиной; самое положительное то, что она, несмотря на болезнь, взялась помогать неродному ребёнку);*

7) репрезентация ситуации как серии причин и следствий (*Она давала ему понять, что у неё нет никого и ничего, кроме него, о чём бы она могла заботиться, и как результат она получила непонимание и непослушание со стороны брата*).

Разновидности данных форм РАП варьируют в зависимости от сочетания следующих трёх параметров: объём представленной ситуации, степень осмысления ситуации, степень личного «проникновения» в ситуацию.

Формы РАП, в которых доминирует опора на моральные, нравственные ценности национально-лингвокультурного сообщества, включают в себя следующие разновидности:

1) апелляция к социальным нормам (*такие поступки достойны уважения; она поступила благородно; осуждать её за это нельзя; Варя правильно поступила);*

2) категоризация героев или их поступков по отношению к различным, преимущественно, морально-нравственным, нормам, принятым в данном сообществе, при помощи: а) использования существительных абстрактной семантики, называющих те или иные нравственные качества (*всё, что она делала – самопожертвование)*, выражающих оценку социумом ситуации проявления тех или иных моральных качеств (*это – жертва; она – героиня; Варя – эгоистка и собственница; она – молодец!)*; б) адъективно-субстантивных сочетаний, в которых прилагательные имеют ярко выраженный языковой аргументативный потенциал (*это* ***настоящая*** *сестринская любовь; Варя –* ***самоотверженная*** *девушка; Варя –* ***полная*** *дура);*

в) наречий *(она* ***по-настоящему*** *любила его; Варя поступила* ***мужественно****; Игорь* ***по-глупому*** *поступил);*

г) гиперонимов *человек, девушка,* сопровождаемых определениями различного характера *(она –* ***девушка****, у кото-*

034-Kolmogorova:_Layout 1 18.05.2008 15:47 Page 225

*рой нелёгкая жизнь; Варя – **человек** с большим сердцем; она - светлой и доброй души **человек**);*
3) сравнение с другими, с эталоном (*далеко не каждый как она…; не каждый решится на такое; не каждый способен на такое; в нашем сегодняшнем мире очень мало таких людей, как Варя);*
4) упоминание прецедентной личности, прецедентной ситуации (*она похожа на Соню Мармеладову; в православии человек верующий должен всегда помогать людям, даже если они не отвечают взаимностью*).

Формы РАП, отражающие индивидуальные психологические особенности речевой личности аргументатора, различаются, прежде всего, по типу доминирующей разновидности речевых актов: а) **экспрессивы**: Похвала (*Она не посмотрела на болезнь*! *Она – молодец*!); Обвинение (*В том, что Игорь был несчастен, виноват только он*…); Пожелание (*чтобы таких людей было больше);* б) **репрезентативы**: Предположение (*а могла бы сказать, что болеет; такой ситуации не произошло бы, поведи себя Варя по-другому; брат сам бы потом осознал, что натворил);* Утверждение (*каждый человек имеет право на выбор, и никто не может переубедить его и заставить переменить решение);* Свидетельствования *(я и сама была в такой ситуации, и знаю, что это такое);* в) **директивы***:* Совет, Рекомендация *(нельзя думать всегда «вот он бедненький!»; ей не стоит продолжать также заботиться о нём; ей не следует приносить себя в жертву ради брата; надо и на себя иметь время)*.

Среди форм РАП, обусловленных определённым уровнем владения аргументатором языковыми средствами выражения, выделяются:

1) использование стилистических средств, например, риторического вопроса (*Но зачем было брать чужого ребёнка? Но стоило ли отказываться от своей жизни?),* обращения и «объединяющего» *мы* (*А что дальше? Но Вы, наверное, сами знаете?; мы этого не наблюдаем*), эпитетов (*слепая любовь, золотое сердце*), противопоставлений (*она всё отдала, а он лишь пользовался заботой);*

2) использование аргументативных маркеров (*но…; к сожалению*…, *хотя, всё- таки; ведь* и т.д.).

1.2.2. Особенности эволюции форм речевого аргументативного поведения в процессе становления коммуникативной компетенции старших школьников и студентов

Результаты использования методик статистического и сопоставительного анализа позволили нам получить подтверждение гипотезы о том, что количество и качество форм РАП подвержено эволюционным изменениям в процессе становления коммуникативной и дискурсивной компетенций членов данного лингвокультурного сообщества.

Можно констатировать, что в процессе становления языковой/речевой личности доля форм РАП, опирающихся на исходную ситуацию, текст, уменьшается: девочками-подростками (14-15 лет) употребляются все 7 разновидностей форм, использующих в качестве опоры исходный текст, в возрастной группе 16-17 лет - 6 таких разновидностей, а в группе девушек 20-24 года - 4. В процентном соотношении их доля в первой возрастной группе составляет 45 % от общего числа использованных форм речевого аргументативного поведения, во второй – 28,5 %, в третьей – 29, 3 %. Формирование же навыков речевого аргументативного поведения с опорой на морально-нравственные нормы социума уже завершается к 14-15 годам, о чём свидетельствует стабильное использование всех 4 выявленных нами форм данной разновидности в приблизительно одинаковых пропорциях: 1 группа – 21 %, 2 группа – 25 %, 3 группа – 24,5 %. Что касается тех форм РАП, которые отражают индивидуальные психологические особенности речевой личности аргументатора, то при всём разнообразии используемых типов речевых актов, в каждой возрастной группе можно выделить доминантный: в 1 группе – директивы Совет-Рекомендация, Поучение; во 2 группе репрезентатив Утверждение, в 3 группе - все выделенные разновидности РА используются практически в равных пропорциях. Отметим, что такие разновидности форм речевого аргументативного поведения, как стилистические средства, практически не используются в речи представителей самой младшей возрастной группы (14-15 лет), а в речи представителей двух других групп – используются достаточно широко. Важно отметить, что во второй возрастной группе (16-17 лет) наблюдается ярко выраженный скачок в использовании аргументативных маркеров на фоне их редкого использования в младшей возрастной группе и спада употребления в старшей возрастной группе: в 1 группе испытуемых их количество составило 34 % от общего числа использованных информаторами форм, во 2 – 47,5 %, в 3 – 33,5 %. Можно предположить, что именно в возрасте 16-17 лет имеет место активное овладение языковыми личностями арсеналом и навыками употребления слов-аргументаторов, а в возрасте 16-17 лет завершается овладение стилистическими языковыми средствами.

034-Kolmogorova:_Layout 1 18.05.2008 15:47 Page 226

1.3. Заключение

Результаты проведённого исследования позволяют сделать несколько выводов:

1) речевое аргументативное поведение можно рассматривать как один из жанров общения (речевых жанров), поскольку для его реализации представители определённого национально-лингво-культурного сообщества используют достаточно стабильный репертуар речевых форм;

2) доминирующим когнитивным базисом разновидностей форм РАП выступают такие виды обобщения опыта, как представления: представления об окружающей среде, «реальности» (представления аргументатора об исходной ситуации), социально-ценностные представления, индивидуально-личностные представления (представление говорящего об уместности и желательности того или иного вида речевого акта в ситуации аргументации); в то время как знания (даже процедурные знания, так называемые «знания как») являются маргинальным видом обобщения опыта, используемым говорящими в аргументативном поведении (знание тех или иных средств языкового выражения);

3) диапазон и частотность форм реализации речевого жанра аргументации варьируют в зависимости от возрастной категории языковых личностей и соотносятся с некоторыми общими тенденциями психического развития и становления личности.

Таким образом, **аргументация**, будучи рассмотрена как один из аспектов «языкового существования» (Гаспаров 1998), речевой повседневности (Колмогорова 2008), должна быть определена не как вербально выраженная рациональная часть убеждения (Рузавин 1997), но **как речевой жанр, вид речевого поведения, разновидности форм реализации которого осваиваются языковыми личностями в онтогенезе в процессе накопления опыта речевого и, тесно переплетённого с последним, социального, взаимодействия в рамках определённого лингвокультурного сообщества и являются частью коммуникативной компетенции представителей данного сообщества**.

Список литературы

1. Бахтин М.М. Марксизм и философия языка // Театрология. М.: Лабиринт, 1998. С.298-456.

2. Борисова И.Н. Непрямая коммуникация в речевой систематике // Прямая и непрямая коммуникация: Сб.науч.статей. Саратов: Изд-во ГосУНС «Колледж», 2003. С.60-71.

3. Выготский Л.С. Мышление и речь // М.: Лабиринт, 2007.

4. Гаспаров Б.М. Язык. Память. Образ. Лингвистика языкового существования // М., 1996.

5. Колмогорова А.В. Неолингвистика или лингвистика повседневности // Проблемы и перспективы языкового образования в XXI в. Новокузнецк: Изд-во КузГПА, 2008.С. 61-68.

6. Колмогорова А.В. Языковое значение и речевой смысл: функционально-семиологическое исследование прилагательных-обозначений светлого и тёмного в современных русском и французском языках // Дисс. … докт. филолог. наук. Кемерово, 2006.

7. Красных В.В. «Свой» среди «чужих»: миф или реальность? // М.: Гнозис, 2003.

8. Леонтьев А.А. Основы психолингвистики // М.: Смысл, 1997.

9. Матвеева Г.Г. Скрытые грамматические значения и идентификация грамматического лица («портрета») говорящего // Дисс. … докт. филолог. наук. С.-П., 1992.

10. Матурана У. Онтология наблюдения // www.philosophy.ru, 2002.

11. Рузавин Г.И. Логика и аргументация // М.: Юнити, 1997.

12. Седов К.Ф. Дискурс и личность // М.: Лабиринт,2004.

13. Соссюр де Ф. Курс общей лингвистики // М., 1998.

14. Холодная М.А. Психология интеллекта: парадоксы исследования. Москва-Томск, 1997.

15. Щерба Л.В. Языковая система и речевая деятельность // Л.: Наука, 1974.

035-Komarova_A:_Layout 1 18.05.2008 15:26 Page 227

ПАУЗАЦИЯ В ЯПОНСКОМ ЯЗЫКЕ НА ГРАНИЦАХ СИНТАКСИЧЕСКИХ ЕДИНИЦ РАЗНОГО УРОВНЯ: КОРПУСНОЕ ИССЛЕДОВАНИЕ[1]

PAUSES ON THE DIFFERENT TYPES OF SYNTACTIC BOUNDARIES IN JAPANESE: A CORPUS STUDY

Комарова А.Д. *(komarovochka@gmail.com)*
Российский государственный гуманитарный университет

Данное исследование посвящено паузам на границе синтаксических единиц разного уровня в устной японской монологической речи. Его цель состоит в том, чтобы проследить, насколько частотны (и вероятны) паузы на границах предложений и меньших синтаксических единиц (клауз), а также какова их «нормальная» длина.

Данная работа представляет собой корпусное исследование на материале устных монологических рассказов на японском языке.

Ее цель заключается в том, чтобы проследить, с какой частотой в устной японской речи возникают паузы на границах синтаксических единиц разного уровня и какова может быть стандартная длина этих пауз. Кроме того, для каждого типа границ прослеживается соотношение абсолютных пауз и пауз заполненных (как лексически, то есть, элементами, имеющими словарное толкование типа русского «ну», так и долексически, то есть единицами типа «мм»), так как заполнение пауз часто может указывать на затруднения говорящего при порождении текста.

Для анализа использовался корпус из тридцати трех текстов, «рассказов по картинкам». От каждого информанта записывалось по четыре текста, однако от двоих инфомантов из-за плохого качества записи использовались неполные комплекты (по три и два текста). Тексты записывались в два приема. Сначала записывались рассказы, а через 6-8 часов – пересказы. Носителям предлагались два набора картинок с несложным сюжетом. Просмотрев одну серию картинок, информанты составляли по ней рассказ, имея картинки перед глазами. Информанты были предупреждены, что эксперимент состоит из двух частей, но им не сообщалось, в чем именно заключается вторая часть. При записи пересказов картинок они не видели и восстанавливали сюжет по памяти.

Паузация представляет собой неотъемлемую часть устного дискурса, однако долгое время при изучении языка за основу бралась письменная форма, и такие распространенные в устном дискурсе явления, как, например, заполненные паузы, ассоциировались только с хезитацией и нарушениями плавности речи.

Во второй половине двадцатого века появились работы, рассматривающие паузацию в связи с синтаксическими границами. Так, в [Chafe 1994] вводится понятие «интонационных единиц» и показывается, что в 60% случаев в английском языке они соответствуют клаузам.

В [Кривнова, Чардин 1999] описывается паузирование на границах «интонационных фраз» в русском языке. Авторы отмечают, что паузы используются для передачи синтаксических и смысловых отношений. Характерно, что большинство границ предложений в русском языке отмечается паузами, и паузы часто совпадают с границами простых предложений в составе сложного.

Данные японского языка по устному дискурсу представлены не столь хорошо, как русского и английского, однако существует ряд работ, посвященных этой теме. Так, [Iwasaki 1993] отмечает, что интонационные единицы в японском языке, как правило, мельче, чем в английском. Однако, [Sakura, Fuji 2006] показали, что большинство интонационных единиц соответствуют клаузам, а исключение составляют топики, более характерные для японского, чем для английского языка.

Данное исследование посвящено паузации на границах синтаксических единиц двух типов в японском языке:

1. Границы предложений

2. Границы между клаузами, не являющиеся границами предложений (т.е. в общем случае границы между простыми предложениями в составе сложного)

[1]Работа выполнена при финансовой поддержке РФФИ, грант 07-06-00061а

035-Komarova_A:_Layout 1 18.05.2008 15:26 Page 228

Сегментирование устного дискурса – одна из важных проблем, возникающих при транскрибировании. Если на письме автор текста как правило использует пунктуационные знаки для обозначения синтаксических границ, то в устном дискурсе подобных средств нет. Вместо графических символов используются просодические характеристики: так, на наше понимание текста влияют интонация и наличие пауз.

Как и письменный, устный дискурс членится на предложения, а также меньшие предикации в составе сложных предложений. Сегментирование текста не всегда тривиально, поэтому в данной работе будут рассматриваться только стопроцентные границы клауз. Такие границы выделялись на основании синтаксического критерия.

В первую очередь, это границы между предложениями. Как правило, конец предложения маркируется адрессивными формами глагола на *masu/desu*. Большинство предложений корпуса оканчивается глаголом в адрессивной форме. Всего в корпусе 301 предложение. Из них 227, т.е. 75,4%, оканчивается на адрессивные формы глагола. Вот пример такого предложения:

(1) DAI T-2

1.	***Kyou***	***wa***	***…(0.8)***	***e(0.7)***	***oyasumi***	***des-u.***
	сегодня	*TOP*			*HON-выходной день*	*COP.ADR-PRS*

Сегодня выходной день.

Исключение составляют заголовки (пример (2)) и подытоживающие фразы с семантикой «конец», где может не быть финитного глагола. Всего таких случаев в корпусе 10 (что составляет 3,3 %).Новым также считается предложение, начатое с обрыва (см пример (7)). В корпусе всего 12 обрывов (т.е. 4% от числа предложений).

(2) WAK R-1

1.	***Nimaime***	***no***	***ue,***	***...(0.6)***	***dainikkaime.***
	второй лист	*GEN*	*согласно*		*второй раз*

Второй лист, второй раз.

Отдельную проблему представляет собой вычленение предложений, которые оканчиваются на неадрессивные формы глагола. Так, в японском языке глагол, употребленный в «простой», т.е. финитной неадрессивной форме, может употребляться как сказуемое или как определение. Часто симптомом конца предложения в корпусе может являться частица, следующая после простой формы, как в примере (3):

(3) ISI T-1

6.	***...(0.9)***	***Baggu***	***..(0.3)***	***ranpu***	***…(0.7)***	***ee(0.4)to***	***nabe***	***ni***
		сумка		*лампа*		*HEZ*	*кастрюля*	*DAT*
	..(0.2)	***kabin***	***iroiro***	***na***	***mono***	***ga***	***ar-u***	***na(0.2).***
		ваза	*разный*	*ATR*	*вещь*	*NOM*	*иметься-PRS*	*PRT*

Есть разные вещи: сумки, лампы, кастрюли, вазы.

В целом же только один из рассказчиков предпочел для двух из своих текстов простые формы и их функция очевидна из контекста. Вот такой отрывок:

(4) ISI T-1

9.	***kime-rare-na-i***					
	решать-POT-NEG-PRS					
	не могу решить.					
10.	***Soo***	***da.***				
	так	*COP.PRS*				
	Так вот.					
11.	***soko***	***ni***	***kodomo-tachi***	***ga***	***i-ru***	***no***
	там	*DAT*	*ребёнок-PL*	*NOM*	*быть-PRS*	*NML*
kii-te			***mi-yoo***			
спрашивать-CNV			*AUX.M-HOR*			

Попробую спросить у детей, которые здесь есть.

В 9-ой строке употреблена простая форма, однако она не может быть понята, как определение к 10-ой строке, поскольку 10-ая строка не начинается с именной группы. То же относится к 11-й строке.

Не сигнализируют о границе предложения адрессивные формы глагола, употребленные с противительным союзом *ga*:

(5) WAK T-1

21.	***....(2.2)***	***Totemo***	***yosasoo***	***na***	***kuruma ga***	***ar-imash-ita***	***ga,***
		очень	*хороший*	*ATR*	*машина NOM*	*иметься-ADR-PST*	*но*

035-Komarova_A:_Layout 1 18.05.2008 15:26 Page 229

22.	***….(0.6)***	***nedan***	***o***	***kik-u***	***to,***
		цена	*ACC*	*спросить-PRS*	*TEMP*

Хотя и была очень хорошая машина, (когда) спросил цену…

Есть также 1 пример неканонического конца предложения:

(6) SHU T-1

8.	***o-too-san***	***anmashi***	***o-kane***	***ga***	***na-kute***
	HON-отец-сан	*почти*	*HON-деньги*	*NOM*	*иметь.NEG-CNV*
9.	***..(0.2)***	***ka-e-na-i***	***to***		
		купить-POT-NEG-PRS	*выходит так*		

(так как) у отца почти нет денег, выходит так, что купить он не может.

В данном случае информанты утверждают, что это конец предложения. Возможно, что финитный глагол после союза (видимо, цитационного) *to* опущен.

Кроме границ предложений, «стопроцентными» границами клауз можно считать границы после:

1. конвербов на *te/de* (не в составе аналитических форм глагола)
2. конвербов на *i* (совпадающих со второй основой глагола и так же не в составе аналитических форм глагола)
3. союзов (условно-временного *to*, соединительного *shi*, противительных *ga*, *keredo(mo)*, уступительного *noni*, причинного *node* и т. п.); союзных имен (например, *toki*) и следующих за ними падежных послелогов (таких, как *ni*) и частиц (например, *wa*); послелогов в составе предикатной группы (*ni, kara, made* и др). А также граница клауз проводится ПЕРЕД цитационным союзом *to* с глаголами мышления и говорения (e.g. *to ui, to kangaeru, to omou*…) Наряду с цитационным союзом рассматриваются цитационные конструкции *to/sou iu fuu ni*

1. Паузы на границах предложений

Итак, по данным корпуса на границе предложений почти всегда - т.е. почти в 94 % случаев - возникают паузы. Из них в 23% случаев это заполненные паузы.

Вот редкий случай, когда паузы на границе предложений нет:

(7) KEN T-1

38.	***….(9.2)***	***Sorede(0.3)***	***sooyu=***	***sooyu***	***koto***	***o***	***==***
		тогда		*вышеназванный*	*NML*	*ACC*	
39.	***Uchi***	***e***	***modot-te***		***kara(0.3),***		
	дом	*в*	*возвращаться-CNV*		*после*		

после возвращения домой…

В приведенном примере видно, что говорящему не удается породить то предложение, которое устроило бы его, и поэтому он бросает неудачный сегмент и начинает новый. При этом времени на планирование у него довольно много: перед забракованным отрывком он делает длительную паузу. К тому моменту, когда он решает отменить неудавшийся отрывок и начать предложение заново, он, видимо, уже знает, что именно хочет сказать.

Вот еще один пример, когда в начале предложения нет паузы (см. начало 3-й строки):

(8) MAY T-2

2.	***..(0.3)***	***asagohan***	***o***	***tabe-mash-ita.***				
		завтрак	*ACC*	*есть-ADR-PST*				
3.	***Sono***	***ato***	***ni***	***…(0.8)***	***ee(0.6)to***	***sukii***	***ni***	***dekake-mash-ita.***
	это	*после*	*DAT*		*HEZ*	*лыжи*	*DAT*	*отправляться-ADR-PST*

Позавтракал. После этого отправился кататься на лыжах.

В данном случае говорящий, предположительно, заранее спланировал несложный отрезок. Второе предложение совсем короткое и, в принципе, говорящий мог бы слить эти два предложения в одно. Кроме того, второе предложение начинается союзным наречием, которое представляет собой эксплицитный показатель связи и также могло повлиять на длину (в данном случае - отсутствие) паузы.

В целом же можно говорить о том, что обычно в позиции границы предложения паузы возникают. Эти паузы дают говорящему возможность сделать вдох и спланировать следующий отрывок, а слушающему наряду с другими просодическими характеристиками служат сигналом о наличии синтаксической (как правило, обусловленной семантическими связями) границы.

Что касается «нормальной» длины пауз на границе предложений, то это очень нестрогий параметр. Длина пауз может варьироваться в зависимости от темпа речи говорящего, жанра текста, более или менее знакомой говорящему тематики и т.д.

Среднее арифметическое не может считаться вполне надежным показателем, т.к. одна слишком затянутая пауза или, напротив, несколько коротких предложений с незначительными паузами на границе могут сильно повлиять на это значение.

035-Komarova_A:_Layout 1 18.05.2008 15:26 Page 230

Поэтому для определения «стандартной» длины пауз на границах предложений в устном неподготовленном рассказе использовались, кроме среднего арифметического, такие показатели, как мода и медиана. Безусловно, сами эти показатели варьируются в разных текстах, однако среднее значение медианы для длины пауз на границе предложений в данном корпусе составило 1,4 секунды, и самая частотная длина паузы на границе предложений в корпусе (т.е. мода) совпадает с этим показателем.

2. Паузы на границах клауз

2.1. Паузы после конвербов на te/de

Конвербы на *te/de* – наиболее частотная инфинитная форма глагола, которая встречается в японской речи. Всего в корпусе 136 клауз, заканчивающихся на данную форму. И приблизительно в 55% случаев на таких границах возникают паузы. Из них в 15 % случаев эти паузы – заполненные. При этом данный показатель сильно варьируется для разных текстов (тенденция может соблюдаться для говорящего в целом): так, говорящий может вообще не делать пауз после форм на *te/de*, как в тексте KEN T-2 или, напротив, отграничивать паузой каждую клаузу на *te/de*, как в тексте MAY T-1.

Вот пример границы клаузы на *te/de* без паузы:

(9) WAK R-2

7.	***..(0.2)***	***sukii***	***o***	***hai-te,***
		лыжи	*ACC*	*надеть-CNV*
8.	***dokoka***	***e***	***dekake-te***	***ik-imash-ita***
	куда-то	*в*	*выйти-CNV*	*AUX.DIR-ADR-PST*

Надев лыжи, куда-то вышел.

И с паузой:

(10) SHU T-2

7.	***….(1.3)***	***yopparat-ta***	***mama***	***sikii***	***o***	***shi-te***
		пьянеть-PST	*так.как.есть*	*лыжи*	*ACC*	*заниматься-CNV*
8.	***…(0.9)***	***koron-de,***				
		упасть-CNV				

[будучи] пьяным катаясь на лыжах, упал..

Средняя длина пауз после данной формы составляет около 0,5 секунды, медиана - 0,6 секунды, а мода – 0,2 секунды.

2.2. Конверб на –i

Формы на *–i* встречаются в текстах значительно реже, чем формы на *te/de*. Всего в тестах встретилось 24 клаузы, заканчивающихся на данные формы. В 16 из 24 случаев после них были паузы, четверть из них (всего четыре паузы) были заполненными. Таким образом, процент пауз после форм на *–i* несколько выше, чем после форм на *te/de,* и составляет 67%, однако выборка слишком мала и не дает оснований утверждать, что паузы после форм на *–i* вообще частотнее.

Что касается таких параметров, как мода, медиана и среднее значение, то они также весьма условны. Медиана и среднее значение очень близки и составляют приблизительно 0,6 секунды. Мода равна 0,4 секунды.

Примеры клауз, оканчивающихся конвербом на *–i*:

(11) KEN T-2

7.	***sukii ita***	***o***	***moch-i,***
	лыжи	*DO*	*брать*
8.	***sukii jou***	***ni***	***dekake-masu.***
	трасса	*ALL*	*выходить-ADDR.PRS*

Позавтракав. (он) взял лыжи и вышел на трассу.

(12) MAI R-2

3.	***...(0.5)***	***ee(0.2)***	***choushoku***	***o***	***tab-e,***		
			завтрак	*ACC*	*есть-CNV*		
4	***...(0.5)***	***de***	***sorekara***	***..(0.1)***	***sukii***	***ni***	***dekake-mash-ita.***
		затем	*после.этого*		*лыжи*	*DAT*	*отправляться-ADR-PST*

Позавтракав, отправился кататься на лыжах..

2.3. Паузы на стопроцентных границах клауз, кроме клауз, заканчивающихся конвербами на te/de и –i

Кроме границ клауз, которые оканчиваются на глагольные формы *te/de* и *–i*, в корпусе встретилось еще 159 стопроцентных границ клауз. Это границы после союзов и послелогов в составе предикатной группы, а также границы перед цитационными союзами.

В корпусе встретились следующие показатели:

035-Komarova_A:_Layout 1 18.05.2008 15:26 Page 231

показатель	**вхождений в корпусе**
ato «после»	9
to-TEMP «если/когда»	14
kedo «хотя»	6
node «так как»	24
-tara «если»	5
toka «или»	5
kara «после»	5
ga «но»	25
keredomo «несмотря.на»	3
tameni «благодаря»	1
niyotte «в.зависимости»	1
-tari репрезентатив	1
ni (в том числе *toki ni, sugu ni, mae ni, 2-я основа+ni, 1-я основа+ zu+ni*) DAT с разными оттенками значения	10
toki (+toki wa) «во.время»	2
-shi «и»	2
-nagara «делая»	1
sei de «из-за»	2
mama «как.есть»	1
-zu отрицание	1
-ba «если»	1
-temo «даже если»	1
dakara «так как»	1
to QUOT цитационный союз	36

Таблица 1

Процент пауз на таких границах достаточно высок и составляет в среднем около 63% . Из них в 18% случаев эти паузы являются заполненными.

(13) WAK T-1, 22

22.	***nedan***	***o***	***kik-u***	***to***	
	цена	*ACC*	*спросить-PRS*	*TEMP*	
23.	***totemo***	***taka-i***	***node,***		
	очень	*высокий-PRS*	*так.как*		
24.	***..(0.4)***	***ka-u***	***koto***	***ga***	***deki-mas-en***
		купить-PRS	*NML*	*NOM*	*мочь-ADR-NEG*

(когда) спросил цену, поскольку (она была) очень высокая, купить не смог.

(14) ISI T-2

1.	***…(0.6)***	***tokoroga***		***..(0.4)***	***sukii***	***ga***	***..(0.3)***	***hm***	***…(0.8)*** ***eto***
		но			*лыжи*	*SUBJ*			
	sakamichi		***ni***	***sashikakat-ta***			***toki***	***ni***	
	уклон		*IO*	*зацепиться-PAST*			*время*	*TEMP*	
2.	*..(0.3)*	***toma-ru***			***koto***	***wa***	***deki-zu***		***ni***
		останавливаться-INF			*SBST*	*TOP*	*мочь-NEG*		*FOC*
3.	*..(0.3)*	***Watasi***	***wa***	***..(0.3)***	***aaa(0.5)***	***…(0.6)***	***koron-de***		***shimai-mash-ita.***
		Я	*TOP*				*падать-COP*		*заканчиваться-ADDR-PAST*

Но, когда лыжи приблизились к склону, я не смог их остановить и упал.

В (13) примере иллюстрируется употребление условно-временного союза *to* и причинного *node*. При этом после *to* паузы нет, а после *node* небольшая 0.4 сек.пауза есть. В примере (14) приводятся случаи употребления союзного имени и отрицательной формы глагола. В обоих случаях на границе клауз возникают паузы.

Среднее значение длины пауз в данной позиции составляет около 0,85 секунды. Мода и медиана составляют, соответственно, 0,2 и 0,55 секунды. Таким образом, видно, что в приведенных примерах (13) и (14) длина пауз не превышает «нормальную» длину.

Среди пауз на границе клауз, кроме границ после конвербов на *te/de* и *–i* отдельный интерес представляют случаи с цитационными союзами и конструкциями. Вот несколько таких примеров:

(15) SHU T-1

4.	***Nani***	***o***	***ka-oo***	***ka?***
	что	*ACC*	*купить-HOR*	*Q*

035-Komarova_A:_Layout 1 18.05.2008 15:26 Page 232

5. ***to ..(0.2)*** ***kangae-te*** ***i-mas(0.3)-u.***
QUOT *размышлять-CNV* *AUX.PRG-ADR-PRS*
«*Что купить?» - размышляет.*

(16) WAK T-1

25. ***....(1.9)*** ***Sorenara*** ***minikaa*** ***o*** ***ka-oo***
тогда *игрушечная.машина* *ACC* *купить-HOR*
26 ***..(0.2)*** ***to*** ***omot-te,***
QUOT *думать-CNV*
Тогда игрушечную машину купить, подумав…

(17) KEN T-1

15. ***oshie-te*** ***ager-u(0.2)***
объяснять-CNV *дать-PRS*
16. ***to*** ***i-imash-ita.***
QUOT *сказать-ADR-PST*
…объясним», – сказали.

(18) WAK T-1

18. ***sou*** ***dat-ta***
так *COP-PST*
19 ***..(0.3)*** ***to*** ***omot-te,***
QUOT *думать-CNV*
Мужчина: «Так было», - подумав…

В корпусе, как видно из таблицы 1, всего 36 вхождений цитационных конструкций, поэтому невозможно сделать точных выводов о вероятности пауз перед ними. Тем не менее, из 36 случаев употребления цитационных конструкций лишь в 8, что составляет 22%, возникли паузы. Это самый низкий процент пауз для границы клауз. Заполненных пауз в данной позиции не встретилось. Из приведенных примеров только в (14) есть небольшая пауза перед цитационным союзом.

Интересен пример (15), где пауза возникает не перед цитационным союзом, а непосредственно после него. Видимо, говорящий уже решил использовать цитационную конструкцию, но задумался при выборе конкретного глагола мышления/говорения. Как видно по остальным примерам, такие паузы также очень редки (их всего 2 в корпусе) и обычно глагол следует непосредственно за *to* и составляет с союзом одну просодическую единицу. Кроме этого примера, в корпусе есть только один случай, когда *to* отрывается от глагола:

(19) KEN T-1

11. ***..(0.4)*** ***to(0.3)*** ***…(0.8)*** ***ojisan*** ***wa*** ***tazune-Ø*** ***==***
QUOT *дядя* *TOP* *спрашивать-CNV*
спрашивает дядя…

В (19) примере сразу несколько факторов могли вынудить говорящего сделать паузу после *to*. Во-первых, говорящий тянет сам цитационный союз. Во-вторых, предложение оборвано. Это доказывает, что у говорящего были трудности с порождением данного фрагмента, и это могло повлечт за собой паузу после *to*. Кроме того, это единственный в корпусе пример, когда цитационный союз оторван от «своего» глагола именной группой (подлежащим), что также могло повлиять на тесноту связи между ними.

Средняя продолжительность пауз перед цитационными конструкциями составляет 0,375 секунды, медиана – 0, 35, а мода 0,2 секнды.

Итак, в разных синтаксических позициях вероятность появления пауз и значения, связанные с их длиной, различны. На материале рассмотренных текстов получились такие результаты:

1. На границах предложений паузы возникают в 94% случаев, 23% из них – заполненные, их средняя продолжительность составляет 1,55 секунды, медиана их значений – 1,4 секунды и мода – 1,4 секунды.
2. На границе клауз после формы на *te/de* паузы возникают в 55% случаев, 15% - заполненные, среднее значение их длины составляет 0,5 секунды, медиана их длин – 0,6 и мода - 0, 2 секунды.
3. На границе клауз после формы на *–i* паузы возникают в 67% случаев, 25% - заполненные, их средняя длина равна 0,6 секунды, медиана длин – так же 0,6 секунда, а мода – 0, 4 секунды.
4. На других «стопроцентных» границах клауз паузы возникают в 63% случаев, в 18 % случаев они заполненные, их средняя продолжительность составляет 0,85 секунды, медиана их длин равна 0,55, а мода – 0,2 секунды.
5. На границах клауз перед цитационнам союзом *to* и другими цитационными конструкциями паузы возникют в 22% случаев, процент заполненных пауз равен нулю. Их средняя длина составляет 0,375 секунды, медиана их длин – 0,35 и мода – 0,2 секунды.

Таким образом, видно, что паузы на границе предложений возникают чаще всего и имеют максимальную продолжительность. Это показывает, что, когда говорящий заканчивает текущую иллокуцию, ему требуется

035-Komarova_A:_Layout 1 18.05.2008 15:26 Page 233

некоторое время на то, чтобы обдумать продолжение и начать новое предложение. Конечно же, эта пауза используется так же и для дыхания. Эти данные согласуются с данными существующих исследований по японскому языку (таких, как [Sakura, Fuji 2006]) и с результатами корпусных исследований по другим языкам (ср. [Chafe 1994], [Кривнова, Чардин 1999], [Кибрик, Подлесская 2008], [Кибрик, Подлесская 2003]).

На границах клауз, не являющихся границами предложений, паузы также достаточно частотны и возникают более, чем в половине случаев. В среднем их длина почти в два раза меньше длины пауз на границе предложений, медиана и мода их длин значительно меньше, чем у пауз на границе предложений. Это подтверждает предположение о том, что, как правило, планирование текста в большей степени происходит на границах больших синтаксических единиц, а паузы на границах клауз, не равных границам прдложений, могут использоваться говорящим для дыхания, обозначения синтаксических границ и, в меньшей степени, для обдумывания продолжения своего текста. Этот результат также совпадает с результатами предыдущих работ в этой области (см. [Sakura, Fuji 2006]).

Что касается цитационных конструкций, то вероятность и длина пауз на границах клауз перед ними минимальна. Видимо, планируя предложение с цитатой, говорящий уже знает, что это цитату надо будет оформить специальным образом, и ему не требуется дополнительное обдумывание на границе клауз. На это указывает также отсутствие заполненных пауз перед цитационными конструкциями. Цитационные конструкции ранее отдельно не рассматривались на предмет возможности пауз до и внутри них, и в этой области данная работа предоставляет новые данные.

Список литературы

1. Кибрик А.А., Подлесская В.И., К созданию корпусов устной русской речи: принципы транскрибирования // Научно-техническая информация. Серия 2: Информационные процессы и системы, № 10. Москва, 2003, 5-11.

2. Кибрик А.А., Подлесская В.И. (ред.) Рассказы о сновидениях: корпусное исследование устного русского дискурса. Москва, 2008. (в печати)

3. Кривнова О.Ф., Чардин И.С. Паузирование при автоматическом синтезе речи // Теория и практика речевых исследований (АРСО-99). Москва, 1999.

4. Хуршудян В.Г. Средства выражения хезитации в устном армянском дискурсе в типологической перспективе. Диссертация на соискание ученой степени кандидата филологических наук. Москва, 2006.

5. Chafe.W. Discourse, consciousness, and time. Chicago, 1994.

6. Iwasaki S. The Structure of the Intonation Unit in Japanese // Japanese/Korean Linguistics, vol. 3, ed. by Soonja Choi. Stanford, 1993, 39-53

7. Sakura C., Fuji S. Intonation units, information structure, and grammatical constructions in Japanese ans English.//Oral presentations. University of Tokyo, 2006.

036-Korotaev_Podlesskaja:_Layout 1 18.05.2008 15:27 Page 234

ФРАЗОВАЯ АКЦЕНТУАЦИЯ В СЛОЖНЫХ ПРЕДЛОЖЕНИЯХ С ПОСТПОЗИТИВНЫМ ПРИДАТОЧНЫМ В РУССКОМ ЯЗЫКЕ: ОПЫТ ИСПОЛЬЗОВАНИЯ УСТНОГО КОРПУСА С ПРОСОДИЧЕСКОЙ РАЗМЕТКОЙ

PROSODY OF CLAUSE-COMBINING IN RUSSIAN: A CORPUS-BASED CASE-STUDY

Коротаев Н.А. *(n_korotaev@hotmail.com),* ***Подлесская В.И.*** *(podlesskaya@ocrus.ru)*
Российский государственный гуманитарный университет

Движения тона в центральных акцентах главной и постпозитивной зависимой клауз могут (1) совпадать, (2) быть противоположными. Кроме того, (3) главная клауза может произноситься безакцентно. Данные устного корпуса со сплошной просодической разметкой позволяют объяснить выбор одной из указанных трех возможностей в реальном дискурсе.

1. Вводные замечания

Данная работа посвящена изучению интонационного оформления полипредикативных конструкций в живой речи*. Ее главная цель – выявить основные принципы взаимодействия коммуникативно значимых акцентов при построении сложноподчиненных конструкций в устном нарративе.

Предпринятое нами исследование носит корпусный характер: как представляется, только работа с реальными корпусными данными позволяет составить верное представление о дискурсивной значимости тех или иных просодических явлений. В качестве исходного материала мы взяли корпус «Рассказы о сновидениях», включающий в себя 129 рассказов детей в возрасте от 7 до 17 лет. Суммарная продолжительность всех аудиозаписей составляет около 2 часов, общее число словоупотреблений – около 15 тысяч. Подробнее о корпусе см. Kibrik et al. 2002, Кибрик, Подлесская, ред. 2008.

«Рассказы о сновидениях» – корпус сравнительно небольшого объема. Однако мы полагаем, что в нем содержится вполне достаточно данных для первоначального анализа интересующих нас явлений: всего в корпусе зафиксировано около 350 полипредикативных подчинительных конструкций. Кроме того, следует подчеркнуть, что для получения достоверных результатов нам нужен был корпус, снабженный сплошной интонационной разметкой, а этого можно достичь только при экспертной обработке «вручную». Этому условию «Рассказы о сновидениях» отвечают в полной мере.

В существенном большинстве сложноподчиненных конструкций, содержащихся в корпусе, а именно, в 281 случае, зависимая часть находится в постпозиции к главной. Мы сосредоточим свое внимание именно на этих примерах, оставив за рамками данного доклада конструкции с начальным и срединным положением зависимой предикации, поскольку последние демонстрируют значительно меньшее разнообразие как в выборе синтаксического типа зависимого, так и в интонационно-коммуникативном оформлении.

Приводя примеры из корпуса, мы будем придерживаться разработанной на его основе системы транскрипционной записи устного монологического дискурса. Важнейшим понятием этой транскрипционной системы является *элементарная дискурсивная единица* (ЭДЕ). Под ЭДЕ мы понимаем минимальный квант, шаг в порождении дискурса. В транскрипции каждая ЭДЕ записывается в отдельную строку. Прототипическая ЭДЕ, с одной стороны, образует единый произносительный контур, с другой – представляет собой простую предикацию. Именно этот канонический случай реализуется в большинстве сложноподчиненных конструкций, см. пример (1) с объектным зависимым:

* Работа выполнена при финансовой поддержке РФФИ, грант 07-06-00061а.

036-Korotaev_Podlesskaja:_Layout 1 18.05.2008 15:27 Page 235

(1) Z56[1]
4. ...(0.5) /Мне приснился /сон,
5. будто мы с /\мамой поехали в \Аме-ерику.

При помощи значков «/», «\», «-» и их комбинаций иконически отмечаются движения тона на акцентированных словах. Как правило, один из акцентов в ЭДЕ является главным, наиболее значимым в коммуникативной, иллокутивной и дискурсивной структуре рассказа (ср. понятие коммуникативно релевантного акцента в Янко 2001: 36-37). Такие акценты мы называем *несущими* и отмечаем их подчеркиванием ударной гласной соответствующего слова. В примере (1) несущие акценты расположены, соответственно, на словах сон и Аме-ерику.

Правила локализации главного акцента в рамках коммуникативных составляющих, не превышающих по объему простую клаузу, прекрасно описаны в работах Ковтунова 1976, Кодзасов 1996, Янко 2001. Напомним, что место коммуникативно релевантного акцента определяется в соответствии с синтаксическим типом фразовой категории и не зависит от порядка слов или коммуникативного статуса составляющей. Нас интересует не то, какие слова выбираются в качестве носителей коммуникативно релевантных акцентов в отдельных клаузах, а то, каким образом имеющиеся акценты взаимодействуют между собой при построении сложноподчиненной конструкции (существование специальных правил сочетаний акцентов в сложных конструкциях было блестяще продемонстрировано в работе Падучева 1989). В этом отношении наше исследование близко к работам Т.Е. Янко о дискурсивных функциях просодии – см., например, Янко 2007.

Ключевое понятие, которые мы будем использовать для описания взаимодействия коммуникативно релевантных акцентов, – *акцентная схема*. Именно рассмотрению наиболее частотных типов акцентных схем в сложноподчиненных конструкциях с конечным положением зависимой части и посвящена основная часть нашего доклада.

2. Основные акцентные конфигурации в сложноподчиненных конструкциях с постпозицией зависимой части

Если использовать термины, принятые в нашей дискурсивной транскрипции, акцентная схема некоторой сложной конструкции – это набор движений тона в несущих акцентах составляющих конструкцию ЭДЕ. Так, для приведенного выше примера (1) акцентную схему можно описать формулой «Подъем в главной ЭДЕ + падение в зависимой ЭДЕ»[2]. По нашим данным, 258 из 281 конструкции с постпозицией зависимой части, зафиксированной в корпусе, следуют одной из семи основных акцентных конфигураций. Ниже мы рассмотрим каждую из них в отдельности.

Порядок представления акцентных схем подчинен следующей логике. Сначала будут рассмотрены схемы, акценты в которых связаны между собой принципом *адаптации* – один из них является «зеркальным отражением» другого, при этом за встраивание всей конструкции в более широкий дискурсивный контекст отвечает несущий акцент во второй по порядку (т.е., зависимой) ЭДЕ. Далее будет разобран случай, при котором главная предикация произносится без несущего акцента; после этого – акцентные схемы с *параллельным* движением тонов. В каждом из этих трех случаев возможны два варианта – с падением и с подъемом на зависимой предикации; соответствующие схемы образуют функционально близкие пары. Последняя, седьмая акцентная схема, в которой зависимая предикация произносится со «вставочным» падением, подобной пары не имеет и потому занимает в нашей классификации особое место.

Перейдем теперь к более подробному рассмотрению выделенных нами акцентных конфигураций.

---

[1] Перед текстом примера приводится индекс рассказа в корпусе. Цифра в левой колонке указывает на номер ЭДЕ в рассказе. Подробнее об обозначениях, принятых в используемой нами системе дискурсивной транскрипции, — в частности, о принципах расстановки пунктуационных знаков в конце ЭДЕ — см. в работах Кибрик, Подлесская, 2003, ред. 2008.

[2] При определении акцентной конфигурации мы сознательно «огрубляем» некоторые реально наблюдаемые движения тона в акцентированных словах, приводя их к одному из двух прототипов — падению или подъему. Мы полагаем, что отдельные «сложные» тоны (восходяще-нисходящий, нисходяще-восходящий и проч.), так же как и ровный тон, в ряде случаев не несут специальной семантической нагрузки, а просто являются своего рода «аллофонами» падения или подъема. Разумеется, это не относится к сложным тонам с закрепленной семантикой — например, таким, как восходяще-ровный тон, служащий для передачи значений «неполноты информации» и «открытого списка» (см. Янко 2001, 2006). Подробнее об отображении сложных движений тона в используемой нами системе дискурсивной транскрипции см. в Кибрик, Подлесская, ред. 2008.

036-Korotaev_Podlesskaja:_Layout 1 18.05.2008 15:27 Page 236

2.1. Акцентные схемы адаптивного типа

2.1.1. Схема 1: «Подъем в главной ЭДЕ + падение в зависимой ЭДЕ»

Эта схема, которую мы уже иллюстрировали на примере объектной конструкции в (1), встречается в корпусе чаще всего – почти в половине всех случаев (в 112 из 281, см. таблицу 1 в разделе 4). Наиболее характерна она как раз для конструкций с сентенциальным дополнением, где на ее долю приходится практически столько же примеров, как и на долю всех остальных схем вместе взятых. Кроме того, схема 1 является одной из двух наиболее частотных для конструкций с сентенциальным определением, см. пример (2):

(2) Z52

10. ..(0.2) ”(0.2) ..(0.3) делаю= такие /движ<u>е</u>ния,
11. которые /наяву я не \мог<u>у-у</u> –де-елать,

При схеме 1 коммуникативный центр конструкции, указывающий на ее место в локальной структуре рассказа, приходится на зависимую предикацию: именно в ней происходит падение тона, которое выражает полную (как в (1)) или частичную (как в (2)) иллокутивную завершенность[3]. Что же касается несущего акцента в главной предикации, то он – согласно принципу адаптации тонов – реализуется в виде подъема, который сигнализирует одновременно о тематическом характере ЭДЕ и о ее незаключительной позиции в составе иллокуции (ср. Янко 2007 о наиболее распространенном типе выражения дискурсивной незавершенности, когда простое предложение в составе сложного концептуализуется как тема).

2.1.2. Схема 2: «Падение в главной ЭДЕ + подъем в зависимой ЭДЕ»

Данная схема, которую можно рассматривать как своего рода «зеркальное отражение» схемы 1, представлена в корпусе существенно меньшим числом примеров, причем практически исключительно в объектных конструкциях (всего – 19 случаев, из них 16 – в конструкциях с сентенциальным дополнением), см. пример (3):

(3) N46

4. ....(1.8) Когда я \усл<u>ы</u>шала,
5. ...(0.5) что-о /б<u>о</u>мба гремит,
6. ....(1.8) то я /побеж<u>а</u>ла к командиру,
7. и \сказ<u>а</u>ла это.

Как и в примерах (1) - (2), главная предикация выступает в роли тематической составляющей, тогда как связь сложноподчиненной конструкции с внешним контекстом маркируется на зависимой предикации (ЭДЕ 5), однако здесь речь уже не идет о какой бы то ни было завершенности. Напротив, вся объектная конструкция 4-5 обладает ярко выраженной иллокутивной незаконченностью, поскольку целиком входит в качестве препозитивного обстоятельственного придаточного в состав сложного предложения 4-7. Соответственно, в несущем акценте зависимой ЭДЕ 5 происходит подъем, т.е. прототипическое движение тона, отвечающее за семантику «неконца», а несущий акцент в ЭДЕ 4 адаптируется к последующему подъему и потому реализуется в виде падения. Отметим, что это падение функционально не имеет ничего общего с падением в ЭДЕ 7 этого же примера или же в зависимых ЭДЕ в (1) и (2), поскольку оно не связано с выражением иллокутивной завершенности или подготовки к завершению иллокуции. Как указывается в работе Кодзасов 2002, подобная огласовка тематической составляющей «асемантична и выполняет чисто оформительские функции».

Использование акцентной схемы 2 свидетельствует о крайней высокой степени контроля говорящего за речепроизводством: еще только приступая к произнесению первой ЭДЕ, он уже должен держать в голове сложную интонационную структуру из, как минимум, трех составляющих. В целом устному непринужденному дискурсу несвойственно применение столь сложных структур (см., в частности, Земская и др. 1981, Miller, Weinert 1998), поэтому сравнительно редкая встречаемость акцентной схемы 2 в корпусе не должна вызывать удивления.

2.2. Акцентные схемы с отсутствием несущего акцента в главной ЭДЕ

Рассмотренный в разделе 2.1 случай адаптации движения тона в несущем акценте главной предикации к движению тона в несущем акценте зависимой предикации указывает на известную степень семантической и интонационной интеграции составляющих полипредикативной конструкции. В рамках сложной интонационной структуры синтаксически главная предикация оказывается в коммуникативно зависимом положении относительно синтаксически подчиненной предикации.

[3] Различие между полной и частичной завершенностью в этих двух примерах отмечается посредством пунктуационных знаков в конце зависимой ЭДЕ: полная завершенность обозначатся точкой, частичная – запятой при падении в несущем акценте.

036-Korotaev_Podlesskaja:_Layout 1 18.05.2008 15:27 Page 237

В ряде случаев процесс интеграции заходит дальше – и тогда главная предикация вовсе лишена собственного несущего акцента, она произносится безударно, в рамках одного интонационного контура с зависимой клаузой. По сути, главная ЭДЕ не имеет в этом случае не только иллокутивного, но и собственного коммуникативного значения. В подобной ситуации на конце безакцентной ЭДЕ не ставится никакого пунктуационного знака – см. транскрипцию следующего примера (строка 5):

(4) N47

4. ...(0.5) Когда я спустился /вниз,
5. ..(0.4) я увидел
6. что \дверь \взломана.

Отсутствие коммуникативно релевантного акцента в главной предикации может сочетаться как с падением, так и с подъемом в несущем акценте зависимой предикации. Пример (4) относится к акцентной схеме 3: «Отсутствие акцента в главной ЭДЕ + падение в зависимой ЭДЕ». Схему 4, «Отсутствие акцента в главной ЭДЕ + подъем в зависимой ЭДЕ», можно проиллюстрировать на примере первых двух строк следующего отрезка дискурса:

(5) N14

31. ....(1.3) {КАШЕЛЬ 0.4} ....(1.9) и я' ..(0.3) п-понимаю
32. что это /сон,
33. ..(0.1) пытаюсь себя /разбудить,

В сумме на эти две схемы приходится 25 конструкций из корпуса (см. таблицу 1), подавляющее большинство которых относятся к конструкциям с сентенциальным дополнением (таковы и приведенные выше примеры (4)- (5)). В полипредикативных обстоятельственных конструкциях эти акцентные схемы не встречаются вовсе, а в конструкциях с сентенциальным определением – существенно реже, чем в объектных[4]. Один из немногих случаев использования схемы 4 в определительных полипредикативных конструкциях представлен в ЭДЕ 18-19 примера (6):

(6) N63

16. ....(1.5) /Потом ..(0.1) он \говорит:
17. «У меня /нет никакой \маски!
18. Я ...(0.5) не-ее не ээ(0.2) тот
19. эээ(0.6) кто ты /думаешь <вот>.»,
20. и ...(0.5) *так \далее*.

2.3. Акцентные схемы параллельного типа

Если адаптивная огласовка несущих акцентов в главной и зависимой ЭДЕ и, тем более, отсутствие несущего акцента в главной ЭДЕ являются симптомами коммуникативно-интонационной интеграции частей полипредикативной конструкции, то параллельное движение тонов, напротив, сигнализирует о разъединенности, фрагментации входящих в конструкцию предикаций.

2.3.1. Схема 5: «Падение в главной ЭДЕ + падение в зависимой ЭДЕ»

Акцентная схема с двумя падениями, при которой главная предикация формирует самостоятельную рему, встречается в корпусе 61 раз (см. таблицу 1). Наиболее характерна она для конструкций с постпозитивными обстоятельственными придаточными, см. пример (7):

(7) N27

7. ....(3.3) \Она-а' ...(0.9) как будто= появилась из /стекла,
8. ...(0.8) 'и ....(1.1) \исчезла-а,
9. ..(0.4) когда-а' ...(0.7) /упала /на ....(1.4) \Олину \кровать.

В данном примере представлен случай так называемого поэтапного падения: уровень частоты основного тона, который данный говорящий обычно использует для указания на конец иллокуции (170-180 Гц), достигается в ЭДЕ 9, тогда как в ЭДЕ 8 падение происходит только до уровня 220 Гц. Максимальная степень интонационно-коммуникативной фрагментации наблюдается в том случае, когда падение в главной предикации имеет «точечный», иллокутивный характер, см. пример (8) с интонационно отделенным определительным придаточным (в обеих ЭДЕ падение происходит до уровня 240 Гц):

[4] Следующей стадией взаимной интеграции составляющих полипредикативной конструкции является утрата главным предикатом основной части предикативных свойств и превращение его в лексикализованный маркер того или иного, чаще всего, модального значения. В этом случае, разбор которого не входит в цели данного доклада, вся конструкция порождается в рамках одной ЭДЕ. Характерно, что, по нашим данным, чаще всего подобные неделимые комплексы образуются как раз на базе объектных конструкций. Об одном случае грамматикализации объектной сложноподчиненной конструкции в монопредикативный комплекс см. Коротаев 2007.

036-Korotaev_Podlesskaja:_Layout 1 18.05.2008 15:27 Page 238

(8) N46

50. ....(1.3) Тогда-а мой /командир /меня /наградил /золот-той \медалью.

51. ..(0.3) Которая /стоила /двести \долларов.

В случаях, подобных (8), главная и зависимая ЭДЕ имеют практически равноценный статус в дискурсивной структуре рассказа: они в одинаковой степени «отвечают» за внешние связи всей конструкции.

Схема 5 – вторая по частотности в корпусе после схемы 1. Ей следует большинство конструкций с постпозитивными обстоятельственными придаточными и существенная часть определительных конструкций (столько же, сколько и схеме 1). Для конструкций с сентенциальными дополнениями она характерна в меньшей степени.

2.3.2. Схема 6: «Подъем в главной ЭДЕ + подъем в зависимой ЭДЕ»

Акцентная схема 6 реализована в 30 конструкциях корпуса, в том числе, в строках 34-36 следующего примера:

(9) Z51

33. ...(0.9) ”(0.1) \Во-от,

34. ....(1.0) ну я-а /подбегаю к ним-м,

35. ..(0.3) к э-этой /шторе,

36. чтоб её /открыть,

37. ....(1.2) /открываю,

38. ...(0.5) *в комнате /\полумрак,*

39. /\открываю,

40. и передо мной так /висят☺ .

(0.7) ☺семь☺ ☺\–трупов☺.

Как и в случае (8), в (9) связь с внешним контекстом интонационно отмечается на обеих составляющих конструкции, а главная и зависимая предикации описывают равноправные события основной нарративной линии. В данном примере главная предикация осложнена интонационно обособленным постпозитивным уточнением, что только подчеркивает фрагментированный характер всей конструкции. Схема 6 регулярно применяется в тех случаях, когда рассказчик использует стратегию «множественных тем» (см. Янко 2006), нанизывая одну неконечную предикацию на другую, прежде чем достичь завершения иллокуции. Именно эта ситуация наблюдается в (9), где произнесенной с конечной интонацией ЭДЕ 40 предшествуют шесть ЭДЕ с четко выраженной незавершенной огласовкой.

2.4. Акцентная схема 7: «Подъем в главной предикации + вставочное падение в зависимой ЭДЕ»

В последней из выделенных нами акцентных конфигураций незавершенность сложноподчиненной конструкции, указывающая на ее статус в рамках более протяженной иллокуции, маркируется только на главной ЭДЕ, произносимой с подъемом, тогда как зависимая клауза как бы «проваливается» между соседними информационно значимыми сегментами. Нередко ЭДЕ, попадающие в эту позицию (в концепции Т.Е. Янко данная коммуникативная роль называется «атонической темой», см. Янко 2001: 76-79), произносятся в пониженном регистре и/или с повышенной скоростью. В транскрипции эти особенности передаются смещением текста вниз и курсивом соответственно – см. ЭДЕ 9 следующего примера:

(10) N01

7. и /я была в таких \жёлтых таких /туфля́х таких,

8. ..(0.2) с-с такими с \шнурочка↑ми,

9. *мама купи↑ла,*

10. ...(0.8) \вот,

11. и /я пошла по \метрow.

Схема 7 используется в тех случаях, когда в зависимой предикации содержится фоновая, «скобочная» информация, не входящая в основную нарративную линию. Всего таких примеров в нашем корпусе 11 (см. таблицу 1), еще один из них представлен ниже под номером (11) со вставочным падением в ЭДЕ 65:

(11) N21

63. ..(0.3) они мне что-то /говорят,

64. ..(0.3) я им какую-то /колкость на это отвечаю,

65. ...(0.8) так что \они-и ..(0.2) ну замол= ..(0.3) \замолчали,

66. ...(0.5) /вот, 67. а /дальше уже \обрывается.

Любопытно, что в нашем корпусе схема 7 ни разу не используется в конструкциях с сентенциальным

036-Korotaev_Podlesskaja:_Layout 1 18.05.2008 15:27 Page 239

дополнением. Вероятно, этот факт объясняется большой дискурсивной значимостью дополнительных клауз: в стандартном случае в них содержится главная для понимания всей конструкции информация.

3. Конструкции со сложной зависимой частью

Во всех примерах, рассмотренных выше, зависимая часть конструкции состояла из одной предикации, одной ЭДЕ. Соответственно, не вставало вопроса о том, несущий акцент в какой именно ЭДЕ нужно принимать во внимание при определении акцентной схемы конструкции. Однако в существенном проценте случаев (в нашем корпусе их доля составляет примерно 30%) зависимая часть состоит более чем из одной ЭДЕ. Так, в примере (12) сентенциальное дополнение глагола вспомнила выражено посредством двух ЭДЕ:

(12) N26

46. ....(1.7) потом я /вспомнила',
47. ...(0.6) ' что он хотел пойти в /гараж,
48. ..(0.3) что-то \сделать там.

Определяя акцентную схему примеров подобного рода, мы учитывали акцент на коммуникативно главной ЭДЕ зависимого комплекса. В подавляющем большинстве случаев – это последняя ЭДЕ. Так, мы считаем, что пример (12) следует акцентной схеме 1, поскольку главная предикация произносится с подъемом в несущем акценте, а последняя ЭДЕ зависимой части – с падением. Что же касается восходящего тона в ЭДЕ 47, то он не является парой к подъему в ЭДЕ 46 (как было бы при акцентной схеме 6), а адаптирован к падению в ЭДЕ 48, но уже не на уровне всей сложноподчиненной конструкции, а на только на уровне ее зависимой части.

В конструкциях со сложной зависимой частью в целом реализуются те же акцентные схемы, что и в конструкциях с простой зависимой частью. Выше, приводя данные о частоте встречаемости той или иной схемы в корпусе, мы имели в виду суммарное число для обоих этих типов конструкций.

4. Количественные данные

Сделав необходимую оговорку о способе учета конструкций со сложной зависимой частью, мы можем суммировать информацию о частоте встречаемости выделенных нами акцентных схем в корпусе. Обобщенные данные представлены в следующей таблице:

		Внутренняя организация зависимой части		
		Простая	Сложная	ВСЕГО
Адаптивный тип	*Схема 1:* «Подъем + падение»	74	38	112
	Схема 2: «Падение + подъем»	17	2	9
	Всего	91	40	131
Схемы с отсутствием несущего акцента в главной ЭДЕ	*Схема 3:* «Нуль акцента + падение»	12	3	15
	Схема 4: «Нуль акцента + подъем»	9	1	10
	Всего	21	4	25
Параллельный тип	*Схема 5:* «Падение + падение»	45	16	61
	Схема 6: «Подъем + подъем»	15	15	30
	Всего	60	31	91
Схема 7: «Подъем + вставочное падение»		9	2	11
ВСЕГО		181	77	258

Таблица 1. Основные акцентные схемы в сложноподчиненных конструкциях корпуса

Список литературы

1. Земская Е.А., Китайгородская М.В., Ширяев Е.Н. Русская разговорная речь. Общие вопросы. Словообразование. Синтаксис. М.: Наука, 1981.

2. Кибрик А.А., Подлесская В.И. К созданию корпусов устной русской речи: принципы транскрибирования // Научно-техническая информация, серия 2. М., 2003. №10. С. 5-12.

3. Кибрик А.А., Подлесская В.И. (ред.) Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М., 2008 (в печати).

036-Korotaev_Podlesskaja:_Layout 1 18.05.2008 15:27 Page 240

4. Ковтунова И.И. Современный русский язык. Порядок слов и актуальное членение предложения. М.: Просвещение, 1976.

5. Кодзасов С.В. Законы фразовой акцентуации // Просодический строй русской речи. М.: Институт русского языка РАН, 1996. С. 181-204.

6. Кодзасов С.В. Фазовая символика тона // Арутюнова Н.Д. (ред.) Логический анализ языка. Семантика начала и конца. М.: Индрик, 2002.

7. Коротаев Н.А. Сегментация полипредикативных конструкций в корпусах устных текстов: конструкции с эпистемическими предикатами в аудиокорпусе «Рассказов о сновидениях» // Научно-техническая информация, серия 2. М., 2007. №2. С. 30-36.

8. Падучева Е.В. К интонационной транскрипции для предложений произвольной синтаксической сложности // Вопросы кибернетики. Семиотические исследования. М., 1989.

9. Янко Т.Е. Коммуникативные стратегии русской речи. М.: ЯСК, 2001.

10. Янко Т.Е. Интонация связного текста // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции «Диалог 2006» по компьютерной лингвистике и ее приложениям. Бекасово, 2006. С. 591-596.

11. Янко Т.Е. Интонационные стратегии незавершенного текста // Динамические модели. Слово. Предложение. Текст. Сборник в честь Е.В. Падучевой. Ред. Р.И. Розина, Г.И. Кустова. М.: ЯСК, 2007.

12. Kibrik A.A., Podlesskaya V.I., Kal'kova T.M., Litvinenko A.O. Cognitive structure of narrative discourse: the analysis of children's night dream stories // Нариньяни А.С. (ред.) Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара «Диалог-2002». М.: Наука, 2002. Т. 1. С. 635-647.

13. Miller, Jim, Weinert, Regina. Spontaneous spoken language: Syntax and discourse. Oxford: Clarendon Press, 1998.

037-Kotov_A:_Layout 1 18.05.2008 15:28 Page 241

УПРАВЛЕНИЕ ДИНАМИКОЙ РЕЧЕВОГО ПОВЕДЕНИЯ ВИРТУАЛЬНЫХ КОМПЬЮТЕРНЫХ АГЕНТОВ

CONTROLLING DYNAMIC SPEECH BEHAVIOUR OF VIRTUAL COMPUTER AGENTS

Котов А.А. *(kotov@harpia.ru)*
Российский государственный гуманитарный университет

В докладе представлена модель для управления речевым поведением компьютерного агента (героя компьютерной игры, компонента интерфейса или, в перспективе, подвижного робота). Модель позволяет имитировать «динамику настроения» агента, что определяет поведение агента в коммуникации. В частности, модель строит из набора шаблонов высказываний короткие монологи, отражающие «переживания» агента, и обеспечивает переключения в диалоге между несколькими собеседниками.

Виртуальные компьютерные агенты могут представлять из себя простые программные модули, которые имитируют поведение отдельного организма в большом сообществе – и создаются в исследовательских целях [Sloman, 2001]. Такие программные модули могут также обладать интерфейсом и использоваться для создания правдоподобных персонажей в компьютерных играх или даже в детских матерчатых игрушках [Paiva, Chaves et al., 2003]. Простые технологии имитации эмоционального поведения и эмоциональной речи используются для разработки интерфейсов компьютерных программ – компьютерных персонажей, которые развлекают пользователя своим поведением, являются учителями иностранного языка, наставниками по фитнесу или даже помогают социализации замкнутых школьников и детей иммигрантов [Hall, Woods et al., 2005]. Сложное речевое поведение и «адекватные» эмоциональные реакции агента на собственные неудачи или неудачи пользователя несомненно повысят интерес человека к таким агентам и позволят создать новые типы интерфейсов, общение с которыми подобно общению с живым человеком – такие интерфейсы снизят нагрузку на пользователя и избавят его от предварительного обучения работе с интерфейсом. Подобные исследования в этих областях представлены в работах и сборниках [Picard, 2000; Cassel, Sullivan и др., 2000; Andre, Dybkjær и др., 2004; Tao, Tan и др., 2005; Paiva, Prada и др., 2007]. Есть надежда, что в перспективе подобные речевые технологии будут использоваться в бытовых роботах, взаимодействующих с пользователем в реальном мире, распознающих эмоции пользователя и отвечающих эмоциональными высказываниями в неловких ситуациях.

Ранее мы рассматривали модель компьютерного агента, который, попадая в эмоциональную ситуацию, реагирует на неё одним высказыванием [Котов, 2006] – (см. Рис. 1). Это высказывание может выбираться в зависимости от имитируемого темперамента или настроения агента[1] – однако по одному высказыванию пользователь не всегда сможет определить состояние агента, и не во всех случаях такой короткий ответ будет достаточен для коммуникации. Для более адекватного взаимодействия агент должен уметь строить небольшой монолог, соответствующий своему эмоциональному настроению: предлагать подряд несколько высказываний, пусть даже не объединённых сюжетными отношениями. Ещё более интересная задача состоит в том, чтобы динамически менять эмоциональное состояние агента в монологе (см. Рис. 2). Например, после негативного эмоционального события агент может расстраиваться, ругать себя, потом «нападать» на адресата, успокаиваться и, наконец, предлагать обсудить решение возникших трудностей. Если такая замена состояний происходит быстро, она производит комический эффект, что часто используется в мультипликации для создания привлекательных персонажей. Наша задача несколько сходна с задачей мультипликаторов, с той разницей, что смена состояний агента, последовательность и структура его высказываний, а также итоговый сценарий – формируются алгоритмически и могут различаться в зависимости от параметров внешнего воздействия и самого агента. Перспективная задача состоит в том, чтобы создать агента, который сможет правдоподобно себя вести в потенциально бесконечном количестве ситуаций (пусть не слишком интеллектуально, но «понятно» с эмоциональной точки зрения), строя при этом потенциально бесконечное число высказываний.

[1] Далее во всех случаях, когда упоминается эмоциональное состояние агента – речь будет идти о таком эмоциональном состоянии, которые мы пытаемся имитировать с помощью речевого поведения агента и которое, в случае успеха, будет приписано агенту наблюдателем.

037-Kotov_A:_Layout 1 18.05.2008 15:28 Page 242

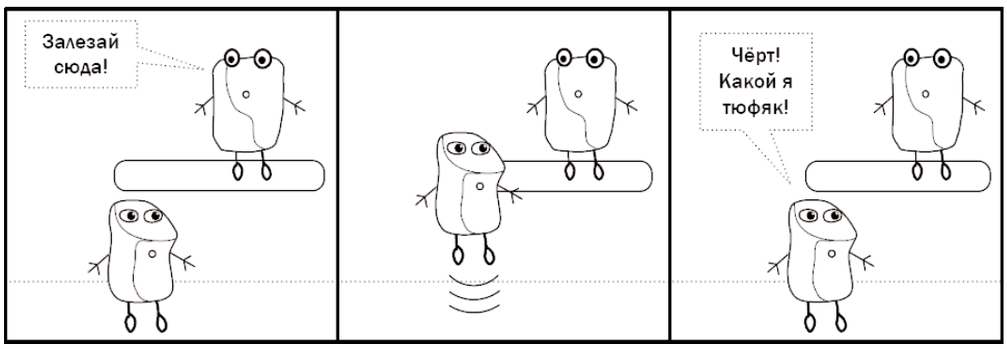


Рис. 1. Базовая ситуация, в которой виртуальный агент переживает неудачу (испытывает фрустрацию)

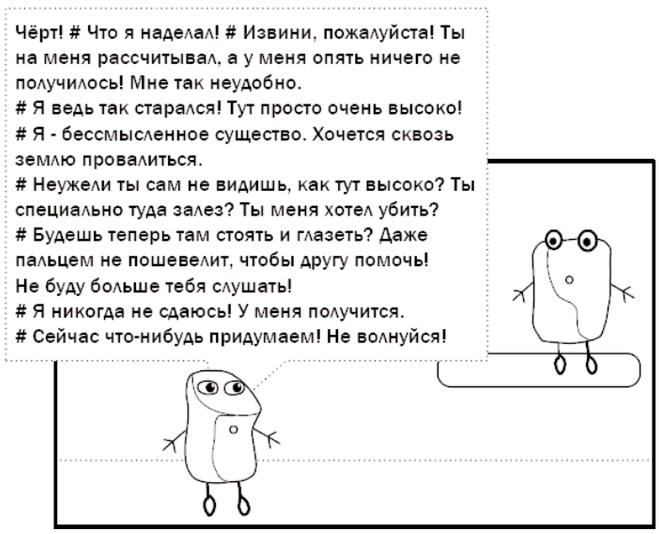


Рис. 2. Смена микросостояний позволяет строить монологи, в которых заметна динамика внутренних состояний агента

В настоящее время предлагаемая нами теоретическая модель реализована в виде достаточно простого компьютерного агента, напоминающего героя мультфильмов или компьютерных игр (интерфейс сделан на технологии Flash). Лингвистическая и поведенческая части агента пока достаточно «негибки»: агент оперирует набором фиксированных фраз (шаблонов) в наборе фиксированных тестовых ситуаций. Вместе с тем, набор используемых фраз достаточно широк и собран на большом экспериментальном материале (около 500 испытуемых разных профессиональных и возрастных групп [Kotov, 2005]), для некоторых фраз предусмотрены альтернативы в выборе лексики для некоторых синтаксических позиций (например, замена местоимений при обращениях к разным участникам ситуации и выбор нейтральной или инвективной лексики – в зависимости от степени «фильтрации»). Хотя такая модель достаточно ограничена, она, тем не менее, хорошо применима для оценки адекватности различных поведенческих стратегий: используемый инвентарь фраз достаточен, чтобы «разыгрывать» разнообразные коммуникативные стратегии и проверять их адекватность с помощью экспертов или в экспериментальном исследовании.

В имеющейся базе размечены соответствия между высказываниями и теоретическими компонентами модели, что открывает возможности для более сложного варьирования высказываний, а также для построения новых высказываний в соответствии с теоретическими правилами.

Архитектура агента

Попытка описать динамику эмоционального поведения агента предпринималась в нескольких научных проектах. В работе [Allen, 2001], посвящённой программной реализации эмоциональных агентов, для этой цели предлагалось использовать набор так называемых «контрольных состояний»: это *планы, намерения, желания, эмоции* (краткосрочные контрольные состояния), а также *предпочтения, навыки, правила* и *личностные характеристики* (долгосрочные контрольные состояния). Предполагается, что конкретное действие агента определяется конкуренцией в иерархии контрольных состояний, каждое из которых обладает разной степенью активизации. В указанной работе эта теоретическая архитектура частично реализована программными средствами для имитации поведения агента (задача имитации речевого поведения при этом не ставится). В нашей

037-Kotov_A:_Layout 1 18.05.2008 15:28 Page 243

работе мы также следуем представлению о том, что выбор высказывания определяются рядом конкурирующих контрольных состояний, например, личностными характеристиками или краткосрочными эмоциями. В частности, мы используем наблюдения над личностными характеристиками разных групп реальных испытуемых, чтобы сформировать и уточнить набор контрольных состояний более низкого уровня – «эмоций» или «настроений».

В другом научном проекте [Becker, Kopp и др., 2004] сделана попытка имитации речевого поведения трехмерного компьютерного агента, причём агент может менять свои эмоции в ходе диалога с пользователем. Речевое поведение агента контролируется тремя общими переменными: pleasure (степень удовольствия или неудовольствия), arousal (степень возбуждения) и dominance (субъективное представление о контроле над ситуацией). Значения этих переменных задают трёхмерное пространство, в котором различные регионы (сферы) отвечают за конкретные эмоции и связанные с ними формы поведения. Например, *страх* определяется как неприятная эмоция, связанная с высоким возбуждением и ощущением потери контроля над ситуацией: [P = -80, A = 80, D = -100]. Входящие стимулы могут менять значения этих переменных (например, оскорбления «расстраивают» агента), что вызывает формы поведения, характерные для нового состояния (агент может выражать недовольство с помощью мимики или даже уходить).

Используемый нами агент обладает двухуровневой структурой. Первой уровень состоит из набора *сценариев* и отвечает за отдельные речевые реакции. Второй уровень состоит из набора *микросостояний* и позволяет агенту отвечать в ситуации сразу множеством высказываний (монологом, не содержащим сюжетных структур) и распределять высказывания во времени, имитируя смену кратких эмоциональных состояний. Представим эти два уровня несколько подробнее.

Простые речевые реакции агента моделируются с помощью набора *сценариев* – правил типа «если – то». Множество сценариев подразделяется на д-сценарии (доминантные сценарии), описывающие речевые или поведенческие эмоциональные реакции, и р-сценарии (рациональные сценарии), описывающие рациональное поведение в эмоциональной ситуации или «рациональные» речевые стратегии, например, ситуация может рассматриваться агентом как задача, требующая решения, или агент может стремится выработать новое правило для поведения в аналогичных ситуациях. Д-сценарии представлены фиксированным набором – 35 единиц[2], в тестовых протоколах используется около 10. Р-сценарии отражают «рациональные» знания агента о ситуациях реального мира или стратегиях этикета; таким образом, для любой ситуации число потенциально возможных р-сценариев бесконечно, но представленные в протоколах ответы можно объединить в группы – всего представлено около 20 классов рациональных ответов. Остальные д-/р-сценарии могут использоваться для синтеза оригинальных или ироничных ответов, которые теоретически возможны, но не представлены в нашей базе (если мы разбили чужую вазу, мы можем сказать: *Как эта ваза красиво шлёпнулась и разлетелась на части!* – однако такой ответ или его варианты не засвидетельствованы в протоколах).

Начальные условия д-сценариев содержат набор стандартных валентностей. Так, негативные д-сценарии включают валентности AGGR (каузатор негативной ситуации) и VICT (жертва негативной ситуации). Если агент произносит фразу *Ты меня чуть не убил*, то такая фраза является следствием активизации д-сценария ОПАСН, в котором агент (говорящий) относит себя к валентности VICT, а адресата – к валентности AGGR. Таким образом, различное распределение валентностей между участниками коммуникации определяет несколько коммуникативных схем. Для негативных д-сценариев это, например, «жалоба», «речевое воздействие», «конфликт», «речевая агрессия», а для позитивных д-сценариев, например, «комплимент», «реклама» и «хвастовство».

Итоговое высказывание, полученное после выбора д-сценария и коммуникативной схемы, подвергается фильтрации. В нашей модели фильтрации подвергаются (а) все высказывания из отдельного класса (например, все фразы, связанные с негативными д-сценариями, могут быть запрещены в этикетном диалоге) и (б) отдельные лексические единицы в составе высказываний – ср. *Куда ты побежал?* (допустимо) *Куда ты попёрся?* (стилистически маркировано и часто некорректно). Фильтрация может не только подавлять некорректные высказывания, но и повышать предпочтение этикетных ответов, именно такая стратегия ответа засвидетельствована в тестовых протоколах абитуриентов (с сравнении с группой «взрослых»).

Степень фильтрации может быть глобальной переменной (агент будет одинаково фильтровать свои высказывания в одной ситуации при обращении к разным собеседникам), или дифференцированной, когда агент будет «официально» говорить с одним присутствующим, и открыто – с другим. В этих случаях используется различный уровень *локальной фильтрации* по отношению к разным участникам ситуации.

Поведение агента во времени и выбор коммуникативных стратегий контролируется несколькими переменными. Среди этих переменных: (а) общее напряжение, сила возбуждения при эмоциональном событии и

[2] Список д-сценариев представлен на странице: http://www.harpia.ru/d-scripts.html

037-Kotov_A:_Layout 1 18.05.2008 15:28 Page 244

скорость угасания возбуждения, (б) предпочтение позитивных или негативных реакций (д-сценариев) – агент может быть «оптимистом» или «пессимистом», предпочтение «рационального» обсуждения ситуации (р-сценариев), (в) предпочтительная коммуникативная схема – агент может любить жаловаться или ругаться, (г) степень фильтрации исходящих высказываний – агент может быть «открытым» или «зажатым», (д) ориентация на себя или на адресата – агент может преимущественно говорить о себе или об адресате (см. Схему 1).

Рис. 3. Фильтрация подавляет некорректные высказывания, и они отображаются в виде «мыслей»

Переключение в коммуникации

Благодаря фильтрации определённых классов высказываний в игровом интерфейсе можно разделить «слова» агента и его «мысли» (высказывания, появляющиеся в «облачке») – это позволяет, к примеру, создать «лицемерного» агента, который на словах ободряет адресата, но сам при этом расстраивается, или наоборот – успокаивает адресата, но сам при этом – злится на него (см. Рис. 3).

Различная степень фильтрации высказываний по отношению к разным участникам ситуации позволяет имитировать переключения в коммуникации, когда агент обращается то к одному, то к другому оппоненту. Если агент испытывает общее напряжение, но при этом из-за требований этикета (высокой локальной фильтрации) не может открыто выразить своё состояние адресату, он может выполнить переключение в коммуникации и обратиться с высказыванием к другому участнику ситуации, для которого уровень локальной фильтрации ниже (см. Рис. 4).

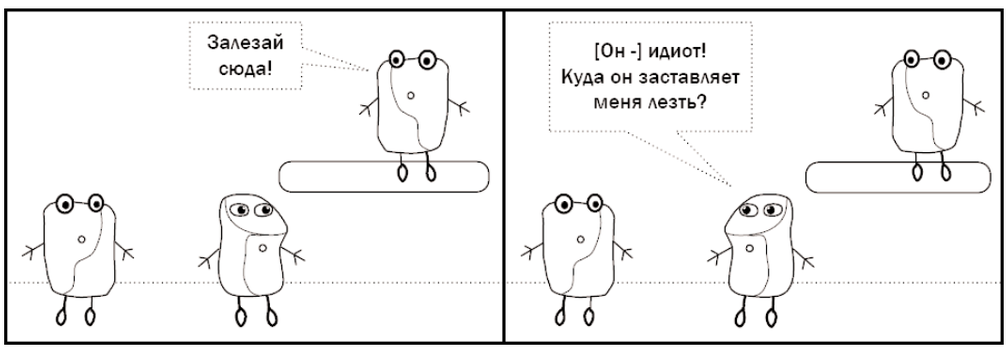


Рис. 4. Переключение в коммуникации: сильный уровень локальной фильтрации заставляет обращаться к другому агенту

Микросостояния

Набор значений общих управляющих переменных, приведённых выше, может быть связан с некоторым коротким состоянием, например, агент может быть сильно возбуждён и на всех ругаться – это состояние будет характеризоваться (а) высоким возбуждением, (б) предпочтением негативных д-сценариев, (в) предпочтением коммуникативной схемы «конфликт» и (г) низкой фильтрацией. Такие краткосрочные состояния, выражающиеся в коммуникации, мы обозначаем термином *микросостояния*. В структуре модели микросостояние определяется как вектор значений управляющих переменных. Каждое микросостояние обладает некоторым уровнем активности – самое активное микросостояние контролирует текущий речевой выход агента. Если микросостояние выражается в речи, то его активизация последовательно снижается, и оно может уступить место другим

037-Kotov_A:_Layout 1 18.05.2008 15:28 Page 245

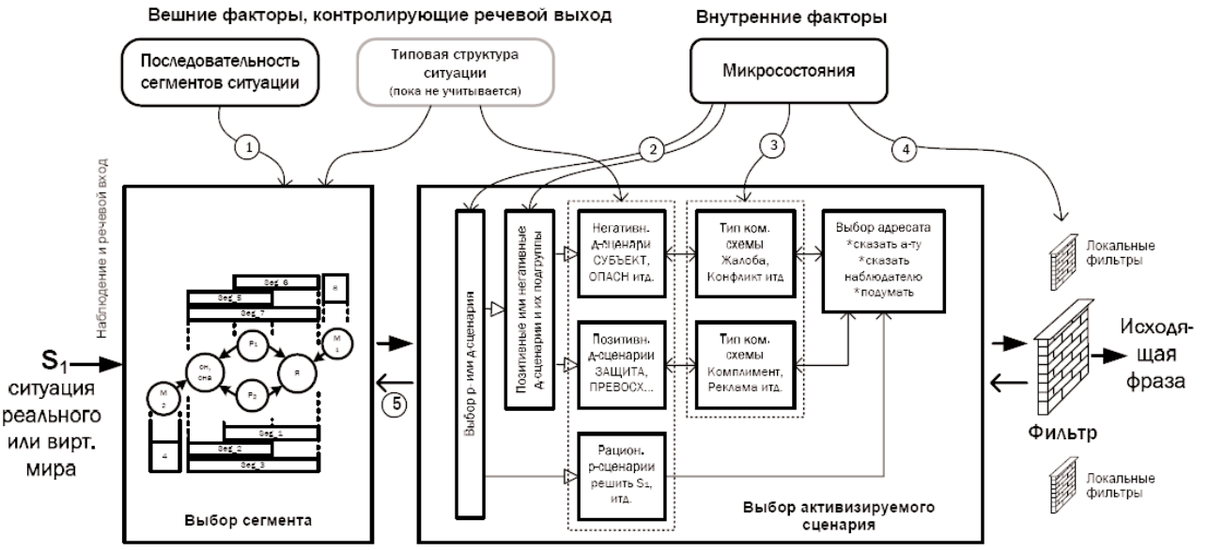


Схема 1. Управление внутренними компонентами модели для имитации динамического речевого поведения

(1) Сегменты ситуации перебираются в определённой последовательности, что создаёт иллюзию «переключения внимания» агента – агент сначала говорит о частных (‘что со мной случилось?’), а затем – о более общих фрагментах ситуации (‘я доставил тебе неприятности’).

Микросостояния определяются как вектор переменных, контролирующих разные компоненты модели: (2) выбор эмоционального или рационального способа реакции, выбор оптимистичной или пессимистичной эмоциональной реакции, (3) распределение участников ситуации в коммуникативной схеме (выбор в ситуации «главного вредителя» или «объекта восхищения») и предпочтение определённой коммуникативной схемы (агент может быть склонен ругаться или жаловаться), и (4) степень глобальной фильтрации высказываний, а также степень локальной фильтрации по отношению к каждому из контрагентов. (5) Предпочтение определённого способа реагирования может влиять на выбор аспекта ситуации (сегмента) – агент может «обращать внимание» на свои действия или на действия других – например, агент может быть «злым», и во всех ситуациях искать свидетельство заговора.

микросостояним. Конкуренция двух микросостояний (последовательное проявление то одного, то другого микросостояния в речи) создаёт эффект эмоционального колебания, например, когда агент то хочет *всё бросить и уйти*, то хочет *проучить адресата*. При подобном колебании снижающаяся активизация микросостояний приводит к выбору всё более «нейтральной» лексики в монологе.

В обозначение микросостояния включаются основные управляющие переменные: указание на сегмент ситуации[3], например, **s5-7** (сегменты 5 и/или 7), **dneg/dpos** — предпочтение негативных или позитивных д-сценариев, **al/ah** — низкое или высокое напряжение (активизация), **fl/fh** — низкая или высокая фильтрация выхода. Нерелевантные для каждого конкретного микросостояния параметры не указываются. Например:

m_dneg — предпочтение негативных д-сценариев: агент эмоционален – ругается или печалится; степень фильтрации или сегменты ситуации не определены;

m_s4-5-7_dneg — агент расстроен/агрессивен и винит адресата (в сегментах 4, 5 и 7 адресат занимает валентность AGGR);

m_s4-5-7_dneg_al_fl — в эмоциональной ситуации агент может флегматично заметить адресату: *Послать*

[3] Сегменты указывают на структуру предиката, который выделяются в текущий ситуации, например, сегмент 1: ‘я сделал P_1’, сегмент 3: ‘я сделал P_1, что затронуло тебя’, сегмент 7: ‘ты сделал P_2, что затронуло меня’. Почти в любой ситуации можно выделить любой сегмент, таким образом, предпочтение сегментов будет в значительной степени зависеть от внутреннего состояния агента, а не от структуры наблюдаемой ситуации.

037-Kotov_A:_Layout 1 18.05.2008 15:28 Page 246

бы тебя куда подальше! В этом микросостоянии агент предпочитает негативные д-сценарии, в которых адресат занимает валентность AGGR, при этом у этих сценариев низкая активизация и низкая фильтрация — мы не сильно переживаем и открыто выражаем наши не слишком сильные эмоции в речи. Данное микросостояние было выделено на основе наблюдений ответов группы практикующих врачей: в эмоциональной ситуации (Вы разбили чужую вазу, женщина-адресат Вас в этом укоряет) более 27% ответов на вопрос *Что бы вы сделали, если бы Вам не нужно было себя сдерживать?* в этой экспериментальной группе имели вид *Послал бы её куда подальше!* (или аналогичные формулировки, относящиеся к тому же классу). В других группах доля ответов этого типа составляла от 2,75% до 6,6%. Из других данных протоколов также следовало, что практикующие врачи демонстрируют достаточно низкий уровень эмоциональной активизации и фильтрации исходящих ответов. Таким образом, эмоциональные высказывания в протоколах врачей вызваны не сильными переживаниями, а низким уровнем фильтрации.

В целом, наблюдение за статистическими различиями тестовых групп позволяет выделить некоторые особенности речевого реагирования. Эти особенности являются неизменными характеристиками речевого реагирования отдельной группы испытуемых, но в нашем случае эта же характеристика может быть преобразована в отдельное микросостояние, которое будет определять специфику поведения агента в течение достаточно короткого времени (2-10 фраз).

Микросостояния могут сменять друг друга, формируя стандартные последовательности. Например, агент может расстроиться и ругаться, но потом – успокоиться и сказать *Ничего страшного* или постыдиться за своё поведение – *Я вёл себя ужасно! Я доставил вам неприятности*. Смена микросостояний происходит за счёт применения ряда правил, микросостояния могут: (а) возбуждать друг друга, (б) подавлять друг друга, (в) одно микросостояние может истощаться и давать место другому микросостоянию.

Экспериментальные исследования

Исследование статистических различий между разными группами испытуемых составляет один из способов выделения и оценки микросостояний. Другой метод – это экспериментальные исследования на группировку эмоциональных высказываний или создание «сценария» из заданного набора эмоциональных фраз.

Мы провели два эксперимента, для которых выбрали по 2-3 характерных примера для каждого из выделенных нами классов высказываний – всего 115 высказываний. Участники каждого эксперимента получали эти высказывания в виде набора карточек.

В первом эксперименте участники должны были разделить высказывания на группы (эксперимент проведен А. Майгуровой под руководством Ю.Е. Кравченко). Для наших целей важны такие протоколы испытуемых, где группировка высказываний осуществляется не по предметному сходству, а по подобию эмоциональных переживаний, связанных с конкретными высказываниями. В этих случаях сходство между протоколами позволяет выделить «центральные» или «ядерные» высказывания для каждой группы, а различия в протоколах позволяют определить «периферийные» или «многозначные» высказывания, которые жёстко не связываются с одним микросостоянием, или которые могут выражать различные микросостояния (например, могут употребляться серьёзно или иронично).

Задача испытуемых во втором эксперименте состояла в том, чтобы сложить из карточек монолог героя мультфильма, который пытается запрыгнуть на верхнюю платформу (Рис.1). Для каждого высказывания испытуемые могли использовать ремарки «подумал» или «сказал», кроме того, испытуемые могли вставлять в свой протокол события «прыгает» (герой делает ещё одну попытку запрыгнуть наверх) или «пауза».

Испытуемые могли написать на протоколе свой адрес электронной почты или оставить протокол анонимным – специальная инструкция разъясняла, что исследование личностных характеристик испытуемых не входит в задачи исследования. В эксперименте приняли участие 30 студентов, 3 протокола были анонимными, на основании остальных протоколов были сделаны мультфильмы, в которых герои разыгрывали представленную испытуемым последовательность фраз и действий. Каждый мультфильм был отослан автору соответствующего протокола для проверки – при этом мы хотели убедиться, что авторы считают свой сценарий адекватным не только на бумаге, но и в форме мультфильма, то есть в ситуации, более приближенной к реальному поведению.

Мы исходили из того, что стандартные группировки и повторяющиеся последовательности фраз в протоколах будут указывать на микросостояния (несколько высказываний, часто упоминаемых вместе, соответствуют одному «внутреннему состоянию») или на регулярные смены микросостояний (близость в тексте фраз, указывающих на «грусть» и «воодушевление» может указывать на то, что «воодушевление» стандартно сменяет выражение в тексте негативных переживаний). Анализ результатов с помощью кластерного анализа позволил в целом подтвердить ранее выделенные микросостояния, а также уточнить степень членства конкретного высказывания в определённом микросостоянии.

037-Kotov_A:_Layout 1 18.05.2008 15:28 Page 247

В целом, проективные и творческие экспериментальные исследования (классификации высказываний или создание сценариев), могут служить дополнительным инструментом для выделения новых микросостояний или их стандартных последовательностей. Кроме этого, как можно предположить, не менее эффективными инструментами могут стать наблюдение за эмоциональной речью и стратегиями поведения в реальных ситуациях, а также анализ стандартных выразительных средств для передачи эмоционального состояния героя в мультипликации и художественной литературе. Эти методы могут позволить создать агентов, демонстрирующих гибкое и правдоподобное эмоциональное речевое поведение при взаимодействии с человеком.

Список литературы

1. Allen S. R. Concern Processing in Autonomous Agents, Ph.D thesis.- University of Birmingham, 2001.

2. Andre E., Dybkjær L., Minker W., и др. (Eds.). Affective Dialogue Systems.- Berlin: Springer-Verlag.-2004.

3. Becker C., Kopp S., Wachsmuth I. Simulating the Emotion Dynamics of a Multimodal Conversational Agent // Affective Dialogue Systems / Ред. E. Andre, L. Dybkjær, W. Minker, P. Heisterkamp.- Berlin, Heidelberg, NewYork: Springer, 2004.- С. 154-165.

4. Cassel J., Sullivan J., Prevost S., и др. (Eds.). Embodied Conversational Agents.- Cambridge, London: MIT Press.- 2000.

5. Hall L., Woods S., Aylett R., Newall L., Paiva A. Achieving empathic engagement through affective interaction with synthetic characters // J. Tao, T. Tan, and R.W. Picard (Eds.) ACII 2005, LNCS 3784.- Berlin, Heidelberg: Springer-Verlag, 2005.- 731-738.

6. Kotov A. A. Application of Psychological Characteristics to D-Script Model for Emotional Speech Processing // J. Tao, T. Tan, and R.W. Picard (Eds.) ACII 2005, LNCS 3784.- Berlin, Heidelberg: Springer-Verlag, 2005.- С. 294-302.

7. Paiva A., Chaves R., Piedade M., и др. SenToy: a tangible interface to control the emotions of a synthetic character // Proceedings of the second international joint conference on Autonomous agents and multiagent systems.- NY: ACM Press, 2003.- С. 1088 - 1089.

8. Paiva A., Prada R., Picard R. W. (Eds.). Affective Computing and Intelligent Interaction. Lecture Notes in Computer Science.- Berlin, Heidelberg: Springer.-2007.

9. Picard R. Affective Computing.- Cambridge, London: MIT Press, 2000.

10. Sloman A. Beyond Shallow Models of Emotion // Cognitive Processing.- 2001.- 2,№ 1.- С. 177-198.

11. Tao J., Tan T., Picard R. (Eds.). Affective Computing and Intelligent Interaction.- Berlin, Heidelberg, New York: Springer.-2005.

12. Котов А. А. Модель эмоционального речевого поведения для виртуального агента ролевой компьютерной игры // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006».- М: РГГУ, 2006.- С. 285-289.

038-Kreidlin_G:_Layout 1 18.05.2008 15:28 Page 248

МЕХАНИЗМЫ ВЗАИМОДЕЙСТВИЯ ВЕРБАЛЬНЫХ И НЕВЕРБАЛЬНЫХ ЕДИНИЦ В ДИАЛОГЕ II Б. ДЕЙКТИЧЕСКИЕ ЖЕСТЫ И РЕЧЕВЫЕ АКТЫ

MECHANISMS OF INTERATION BETWEEN VERBAL AND NOHVERBAL UNITS IN A DIALOG II B. DEICTIC GESTURES AND SPEECH ACTS

Крейдлин Г.Е. *(gekr@iitp.ru)*
Российский государственный гуманитарный университет

Академическая лекция как разновидность диалога служит хорошим экспериментальным полигоном для изучения общих закономерностей и конкретных правил взаимодействия в диалоге речевых и неречевых единиц. В части II А была построена классификация дейктических жестов и охарактеризованы основные типы дейктических лекторских жестов. Ниже, в части II В, показано, что дейктические жесты каждого из типов отличаются своими нетривиальными связями и соотношениями с другими вербальными и невербальными знаками в диалоге.

Ключевые слова: диалог, академическая лекция, дейктический жест, вербальный знак, невербальный знак.

1. Введение

Настоящая работа представляет собой очередной фрагмент большого проекта по описанию некоторых общих механизмов и конкретных особенностей сосуществования в диалоге разных типов вербальных и невербальных знаков.[1] Речь в ней пойдет о коммуникативном взаимодействии с речевыми актами (и с отдельными составляющими этих актов) русских **дейктических (указательных) эмблематических жестов** (или дейктических эмблем) и **дейктических иллюстративных жестов (иллюстраторах)**.

Как и прежде, рассматривается конкретный вид коммуникации – лекторский диалог.[2] Напомню, каковы исходные данные для анализа, его цели и задачи. Изучается вербальное и невербальное знаковое поведение университетских лекторов; реакция студенческой аудитории на произносимые лекторами тексты по предположению должна составить отдельный объект исследования. Все лекторы, их 20 человек, – это люди разного пола, их возраст – от 35 до 65 лет, в основном все они являются штатными преподавателями Российского государственного гуманитарного университета (РГГУ) самых разных гуманитарных профессий: лингвисты, историки, социологи, политологи, специалисты по теории литературы и культурной антропологии. Изучение лекторского поведения проводилось при помощи видеокамеры, на которую записывались части лекций, во время которых шло объяснение нового материала. Впоследствии для анализа отбирались 10 – 15 минут из общего числа записанных. Слушатели, студенты в аудиториях, – это люди обоих полов (впрочем, в силу специфики вуза, преимущественно женского пола), приблизительно одного возраста, обычного темперамента, нормального физического и психического здоровья.

Основной и окончательной целью всего проекта является не столько описание отдельных лекторских, или дидактических, жестов и жестовых[3] комплексов, сколько определение механизмов, установление основных принципов и способов взаимодействия жестового и вербального составов устных высказываний и анализа телесного поведения участников диалога. В этой связи напомню то замечание, которое уже делал в первой части

[1] Настоящая работа выполнена в рамках проекта при поддержке Российского гуманитарного научного фонда (грант РГНФ № 07-04-00203а).

[2] В своих наблюдениях и выводах я опираюсь на визуальные наблюдения и фрагменты видеозаписей лекций, которые читались преподавателями институтов и факультетов в составе Российского государственного гуманитарного университета. Видеозапись и аналитическая обработка материалов проводились мной и студенткой Института европейских культур в составе РГГУ П. Б. Печерской, которая в 2001 году под моим руководством написала и успешно защитила диплом под названием «Семиотика лекторских жестов» (Печерская 2001). Также, с любезного разрешения моей французской стажерки Флоранс Рико-Кассар, я частично пользуюсь собранными ею видеоданными. Она под моим частичным руководством проводит сравнительный анализ невербального поведения французских и русских лекторов, см. об этом в Рико-Кассар 2005.

[3] А также жестово-фонетических комплексов, которые рассматривались в статье Рико-Кассар 2005.

038-Kreidlin_G:_Layout 1 18.05.2008 15:28 Page 249

этой работы, см. Крейдлин 2007. Хотя академическую лекцию обычно рассматривают не как диалог, а как коммуникативно-ориентированный монолог дидактического характера, мне представляется, что такой речевой жанр, как академическая лекция, имеет ряд разновидностей и присущих только им особенностей. В частности, академическую лекцию можно читать по-разному, например, в диалогическом режиме. И мне были интересны именно такие лекции-диалоги, поскольку только в них имеет место явный и постоянный контакт преподавателя с аудиторией.

2. Классы дейктических жестов

Ниже я коротко напомню, о чем по существу шла речь в первой части представляемого здесь исследования (Крейдлин 2007). В этой части были выделены и описаны два семантических класса дейктических жестов – **<собственно> дейктические** и **характеризующие дейктические жесты**.

В толкованиях <собственно> дейктических жестовых лексем основная (и, как правило, единственная) пропозиция выражает идею указания некоторого объекта (объекта в самом широком смысле слова – это может быть предмет, то есть ‘вот X’, его местонахождение ‘X тут’, направление движения ‘X направился туда’ и др.). В смысловом представлении характеризующих дейктических жестов указание не составляет основной пропозиции: оно выполняется только для того, чтобы можно было затем квалифицировать и/или оценить объект или какие-то его свойства. Иными словами, схема толкования этих жестов выглядит следующим образом: ‘указывая на X, сообщаю, что X…‘. Характеризующие дейктические жесты не столько указывают (хотя делают и это тоже) на объект, сколько показывают действия объекта или демонстрируют его свойства. Таким образом, выделенные классы дейктических жестов различаются той ролью, которую в их семантическом представлении играет дейктический компонент.

Подробному анализу были подвергнуты также морфологические и структурные разновидности <собственно дейктических жестов. Форма дейктических жестов определяется значениями трех признаков: (1) ‘каков активный орган жеста и/или какова его рабочая часть’; (2) ‘каково направление этого органа (части) в данном жесте’ и – для мануальных жестов – (3) ‘ориентация ладони’.

Из всех логически допустимых комбинаций значений выделенных признаков в лекционной практике реализуются не все. Встречаются только: (I) **жесты рук** (**рабочий орган – указательный палец**): I.1 указательный палец направлен вертикально вверх/вбок, ладонь ориентирована на жестикулирующего;[4] I.2 указательный палец направлен вертикально вверх/вбок, ладонь ориентирована на адресата; I.3 указательный палец направлен вертикально вниз/вбок; ладонь ориентирована на жестикулирующего;[5] I.4 указательный палец направлен горизонтально вперед/вбок, ладонь ориентирована вниз; (II) **жесты рук** (**рабочий орган – большой палец**): II.1 большой палец направлен вертикально вверх/вбок; II.2 большой палец направлен горизонтально вбок;[6] II.3 большой палец направлен горизонтально назад/вбок; (III) **жесты рук** (**рабочий орган – мизинец**): III.1 мизинец направлен вертикально вверх/вбок; III.2 мизинец направлен вертикально вниз/вбок; III.3 мизинец направлен горизонтально, ладонь ориентирована вниз; (IV) **жесты рук** (**рабочий орган – рука**): IV.1 рука направлена горизонтально вперед, ладонь ориентирована вниз; IV.2 рука направлена горизонтально вперед, ладонь ориентирована вверх; IV.3 рука направлена вертикально вбок/вверх, ладонь ориентирована на адресата; (V) **жесты головы** и **частей головы** (прежде всего – **жесты глаз**).

На множестве указательных жестов регулярно выражаются следующие смысловые противопоставления – большей/меньшей определенности (индивидуализации) объекта и единичности/множественности объекта. Кроме того, еще одно противопоставление относится к характеру целевого объекта, на который направлено указанию. Так, если смысловое задание требует не точной референции объекта, места и т. д., а лишь привлечения внимание к объекту (месту и др.), о котором идет речь либо непосредственно в лекционном материале, либо во внешнем контексте лекции, то используются жесты руки с открытой ладонью вверх или – реже – вниз.

В работе Крейдлин 2007 были описаны также основные смыслы, передаваемые рукой с открытой и расположенной горизонтально ладонью. Она служит носителем смыслов открытости, представления и предоставления адресату объекта, в том числе метафорического, такого, как, например, тема актуального обсуждения в диалоге. Рука с открытой с горизонтальной ладонью открывает адресату возможность иметь дело с данным объектом, вводит объект в личную сферу адресата и как бы предлагает ему/ей взглянуть на объект или

[4] Это жест – достаточно редкий.

[5] Еще одна теоретически допустимая и антропоморфно возможная мануальная жестовая форма – опущенный вниз/вбок указательный палец с ориентацией ладони на адресата – нам практически не встретилась. С ней мы столкнулись только один раз на лекции у преподавателя-женщины.

[6] Очень редко – всего в двух случаях употребления жеста – большой палец был направлен горизонтально вперед.

038-Kreidlin_G:_Layout 1 18.05.2008 15:28 Page 250

«взять» его, а рука с открытой и опущенной вниз ладонью символически придвигает объект ближе к жестикулирующему. Не случайно, этот жест часто используют гиды, ведущие экскурсии по городу и демонстрирующие людям объекты культуры и быта, и экскурсоводы в музеях. Объект при этом метафорически как бы передвигается в личную сферу адресата, и производимый жест говорит о том, что у человека появляется возможность присвоить этот объект или что человек это уже сделал. Одновременно жест свидетельствует о частичной недоступности объекта для адресата жеста (ведь противоположная сторона кисти руки направлена автоматически в сторону адресата!)

3. Дейктические жесты и речевые акты

Ниже описываются особенности взаимодействия отдельных дейктических жестов с разного рода речевыми актами.

Иллюстративные жесты сопровождают разные речевые акты и разные высказывания, и информацию о том, что эти речевые акты или высказывания собой представляют, необходимо, как мы со всей уверенностью полагаем, включать в жестовые словари. И наоборот, в словари речевых актов (или конституирующих их основных единиц) следует включать информацию о том, какие невербальные единицы обязательно или с высокой степенью вероятности сопровождают эти акты. Сказанное относится и к дейктическим иллюстраторам рук и пальцев рук. Так, указательный жест пальцем на адресата речи очень часто сопутствует агрессивным речевым актам, таким как обвинение, упрек, попрек или недоброжелательное замечание, то есть актам, в которых адресат рассматривается как виновник наступления разных негативных событий. Кроме того, эту жестовую форму можно встретить в строгих императивных актах приказа, команды, требования и в языковых высказываниях в модусах обязанности или сильного долженствования (типа «Ты это непременно должен сделать!», «<Не на то обращаете внимание>, Вот о чем надо задуматься!», «Вы будете сейчас выступать в роли критика!»). Поэтому совершаемые на близком – физически личном или физически интимном расстоянии[7] – дейктические жесты пальцами квалифицируются как несанкционированное и негативное, чаще всего грубое и авторитарное, проникновение в личную сферу и во всех известных нам культурах осуждаются.[8]

А вот рука с открытой ладонью, указывая на что-то, одновременно это что-то представляет, предлагает взглянуть или предоставляет, дает что-то адресату речи. И потому для дейктических жестов руки с открытой вверх ладонью провести четкую границу между собственно дейктическим и характеризующим дейктическим жестами очень трудно. В любом случае жесты руки сопровождают совсем другие, чем жесты пальца, акты, а именно вежливые речевые акты со значением предложения, выдвижения некоторой точки зрения, раскрытия содержания отдельных установок, положений, гипотез или аргументов.

Собственно дейктические жесты головой и глазами в лекторских диалогах по нашим видеоматериалам показывает на исписанную доску или на предметы, находящиеся на доске – карты, схемы, макеты, изображения с проектора и др., причем в основном тогда, когда лектор стоит лицом к аудитории. Кроме того, жесты головой и глазами сопровождают когнитивный (но часто воплощаемый в языке) акт отклонения или отбрасывания какой-то мысли, символизируя оставление ее в стороне от магистральной линии повествования. Вместе с тем указания рукой и головой/глазами, как правило, не дают точной локализации целевого объекта.

В процессе чтения лекций дейктические жесты обычно сопровождают высказывания лекторов, которые относятся к классу речевых указаний, что, в общем-то, совсем не удивительно. Интереснее то, что дейктические жесты у лекторов чаще появляются не вместе с речью – исключение составляют абстрактные указания, или дейктические невербальные метафоры, о которых пойдет речь дальше, - а там, где речь по тем или иным причинам отсутствует или ее восприятие затруднено. Жестом лектор выбирает человека в аудитории, когда не знает его или ее имени или фамилии, когда сидящий находится от него далеко, когда не хочет прерывать объяснение материала. Кроме того, дейктические жесты особенно хорошо приспособлены к передаче сведений, касающихся разного рода пространственных отношений, в частности формы и топографии объекта, ориентации и направлении движения и др. признаках, которые тяжелее кодируются вербальным способом. Мониторинг лекторского поведения преподавателей РГГУ показал, что более опытные из них подсознательно или вполне осознанно выбирают невербальные средства, которые наряду с адекватно подобранными вербальными средствами облегчают аудитории понимание и усвоение нового материала.

Ниже я остановлюсь на трех сравнительно часто встречающихся в лекциях преподавателей РГГУ по разным дисциплинам высказываниях и речевых актах, за которыми, так сказать, закреплены отдельные собственно дейктические лекторские жесты, и прокомментирую наиболее интересные из них.

[7] О типах расстояний, используемых в невербальной или вербально-невербальной коммуникации см. подробно в книге Крейдлин 2002.

[8] Грубость указания мизинцем, по-видимому, из-за его небольшой величины, сглажена, и детям такое поведение прощается.

038-Kreidlin_G:_Layout 1 18.05.2008 15:28 Page 251

1. Если материал лекции предполагает указание на высшие силы и, метонимически, на какие-то властные структуры наверху, то используется только жест вертикально направленным указательным пальцем, а жест руки здесь недопустим[9]. *Там всё видят* – произносит преподаватель психологии, демонстрируя данный жест. Направление пальца в этом жесте вертикальное и не наклонное. Ср. также примеры, взятые из художественной литературы: (1) *Когда Сталин в мертвой тишине, подняв указательный палец, говорил: «они там думают», я чувствовал удушье и мне казалось, что и Сталин, и я заживо похоронены в мраморном подвале и избраны нечистой силой охранять труп желтого человечка в кителе с нагрудными карманами, потерявшего последнюю примету отношения к жизни* < ... > (Юз Алешковский. Рука (Повествование палача)); (2) – *Это и были ее последние слова в моей жизни. –- Тсс! –- вдруг сам себя прервал больной и поднял палец, – беспокойная сегодня лунная ночь.* (М. Булгаков. Мастер и Маргарита).

Понятным и естественным выглядит переносное значение жеста, который стал широко использоваться как указание на важные моменты в речевом высказывании, как жестовое средство подчеркивания ключевых когнитивных и психологически важных мест. Поднятый указательный палец призывает обратить внимание на некоторые важные моменты и несет в себе смыслы предупреждения и назидания (что, возможно, связано с семантикой старого и в высшей степени культурно значимого жеста **перст указующий**). Дидактический дейктический жест здесь приобретает функцию характеризующей невербальной единицы, способствующей ритмизации и коммуникативной организации речи, и, посредством апелляции к высшим силам, а также диалогическую функцию невербального знакового средства привлечения внимания слушателей к определенным отрезкам текста. Ср. (3) *Заглянул в микроскоп, радостно и как бы хищно, осклабился. – Я его поймаю, – торжественно и важно сказал он, поднимая палец кверху – поймаю.* (М. Булгаков. Роковые яйца); (4) *– За что это вы его благодарите?- заморгав, осведомился Бездомный. –- За очень важное сведение, которое мне как путешественнику чрезвычайно интересно – многозначительно подняв палец, пояснил заграничный чудак* (М. Булгаков. Мастер и Маргарита).

Быстро поднятый вертикально вверх указательный палец руки как характеризующий дейктический жест имеет еще одно значение иллокутивного акта со значением ‘Вот <оно>!’, ‘Эврика!’, то есть жест имеет значение внезапно найденного или принятого решения. Дейктический жест указательным пальцем руки, поднятой перед головой (отражает то, что найденное решение продуцировано мозгом), перед глазами (отражает то, что решение увидено) и перед носом (экспликация смыслов чувства, проницательности, интуиции), просто предназначен, как утверждает французская исследовательница Ж. Кальбрис, для выражения именно этих смыслов (Калбрис 1990, с. 72). Жест часто сопровождает слова торжества и экспрессии. Слова *Вот, к чему мы пришли*; *Вот он, смысл где* произносят на своих лекциях вместе с жестом, соответственно, лингвист и литературовед.

Дейктические жесты, совершаемые указательным пальцем и употребляемые в разных своих значениях, настолько частотны и узнаваемы, что при их языковых описаниях слово *указательный* обычно опускается. Ср. выражения *не показывай пальцем*, *не тычь в меня палец*, а также описание таких жестов в текстах (5) *Музыка эта возникла скорее всего из-за того, что, как догадались еще древние (тут палец Николая Борисовича указал на полоску ватманской бумаги с черными словами), слух наш по сравнению с другими чувствами куда меньше облагодетельствован естественными наслаждениями* (Вл. Орлов, Альтист Данилов); (6) *Ибо, если бы это было так, ты обязательно взял бы у меня что-нибудь. Имей в виду, что он перед смертью сказал, что он никого не винит, – Пилат значительно поднял палец, лицо Пилата дергалось.* (М. Булгаков, Мастер и Маргарита).

2. Из теоретически возможных дейктических жестов, совершаемых большим пальцем, в лекциях употребляются не все. Лекторам не свойственны жесты, в которых большой палец направлен горизонтально вперед, и жесты с большим пальцем, направленным вертикально вверх. Зато мы часто встречались с жестами больших пальцев, направленными назад. Это, прежде всего, обращение внимания слушателей на то, что написано на доске за спиной лектора, то есть в ситуациях, когда лектор стоит лицом к аудитории; это также горизонтальное указание вбок. Довольно неожиданно для нас у нескольких лекторов жестовое указание большим пальцем имело несколько иное употребление: большой палец метафорически указывал на боковые ветви повествования, на отклонение от его магистральной линии.

Есть еще некоторые особенности дейксиса указательным и большим пальцами. Указательным пальцем показывают вниз, а большим пальцем вниз не показывают. Это обстоятельство определяет особенности сочетания соответствующих форм с речевыми произведениями разной семантики и иллокутивной силы. Указательный палец вниз подчеркивает определенность и окончательность выражаемых суждений. *Вот так именно оно и произошло, началась война* – произнося этот текст, преподаватель истории одновременно резким движением устремляет свой указательный палец вниз. Многократно повторенное движение руки с выпрямленным и устремленным вниз указательным пальцем сопровождает высказывание *И мы им это говорили*

[9] Апелляция и воззвание к Небесам и Господу совершается только поднятыми двумя руками с направленными одна на другую ладонями, но это не дейктический жест.

038-Kreidlin_G:_Layout 1 18.05.2008 15:28 Page 252

на протяжении многих лет. Высказывание это произносит лектор на лекции по политологии. Его жест подчеркивает, что тем, кому это говорилось, «они», – оппоненты, враги, «чужие», – как бы оказались внизу, а говорившие это, с кем солидаризуется также сам лектор, то есть «мы», заняли доминирующую позицию. и могут смотреть теперь на «них» сверху вниз, подчеркивая свое интеллектуальное или иное превосходство. И иконически указательный палец, направленный вниз, именно эту сложившуюся иерархию демонстрирует.

3. Преподаватель итальянского языка необычным жестом, явно относящимся к итальянскому языку жестов, указывает на некоторый пункт высказывания, которое произносит студентка. Впоследствии, отыскав описание этого, известного мне по итальянским фильмам и глубоко заинтересовавшего меня жеста в ряде работ (Де Йорио 1832/2000; Кендон 1994; Поджи 1983; Эфрон 1972), я узнал следующее. Жест по-итальянски обычно называется *mano a borso*, переводится на английский как *purse hand,* что по-русски можно передать как ‘рука в форме кошелька или… щепоти!’.

Mano a borso – это диалогический жест, он встречается в дискурсе как вместе с речью, так и без нее. Относится он, в терминологии А. Кендона, к классу прагматических жестов, то есть является телесным невербальным знаком, сопровождающим конкретный речевой акт либо какие-то структурные или смысловые аспекты высказывания. При воспроизведении жеста пальцы какой-то одной руки, для правшей – правой (жест не исполняется двумя руками), складываются вместе, кончики пальцев касаются один другого. Рука со сложенными пальцами чуть согнута в локте, вытянута горизонтально вперед в сторону адресата, ладонью вверх; она может находиться на разной высоте и на разном расстоянии от средней линии тела жестикулирующего, однако в ряде случаев употреблений жеста рука с тем же положением пальцев направлена на тело самого жестикулирующего.[10] Рука может перемещаться при этом вверх-вниз вдоль тела, совершая повторные движения – все с относительно небольшой амплитудой, однако частота движения может быть разной – более высокая частота означает, в частности, нетерпение или нетерпеливость человека, его высокий темперамент, бóльшую экспрессию поведения. Жест **mano a borso** свойственен скорее мужчинам, или, точнее, мужскому стилю поведения (о гендерных невербальных стилях поведения см. в книге Крейдлин 2005).

Согласно данным, содержащимся в упомянутой ранее литературе, в которой описывается данный жест, он всегда сопровождает вопрос – не обязательно интеррогативное высказывание по форме, но обязательно высказывание с семантикой, или иллокутивной силой, вопроса. И. Поджи считает, что в зависимости от способа исполнения и формы, которую принимает жест **mano a borso**, он может указывать на ту или иную семантическую разновидность вопроса. Жест сопровождает, в частности, обычный вопрос, негативный псевдо-вопрос (что-то вроде *Неужели ты и вправду этого не знаешь? Ну, скажи же!*), вопрос, при котором спрашивающий твердо знает, что адресат знает ответ, но не понимает, почему тот «тянет», медлит с ответом, и еще некоторые другие разновидности вопросов. В лекторской речи, как нам представляется, **mano a borso** применяется для указания на тот или иной момент в вопросе, как бы фокусируя внимание собеседника на этом моменте и подстегивая собеседника к ответу. Именно на такое метафорическое использование жеста в функции дейктического знака обратил внимание А. Кендон в своей работе Кендон 1995, с. 251 – 253.

Заключение

В настоящей работе излагаются некоторые предварительные результаты комплексного исследование особенностей взаимодействия невербальных и вербальных знаков в устном диалоге. На примере лекторских дейктических жестов я попытался показать, что многие из них являются обязательным элементом коммуникативного акта и что жесты с разными формальными и смысловыми характеристиками отличаются своими связями и соотношениями с другими знаками, выступающими в диалоге. Особое внимание было уделено взаимодействию дидактических указательных жестов с речевыми актами и высказываниями разной иллокутивной силы. Отдельные лекторские дейктические жесты сопровождают речевые единицы с настолько высокой степенью обязательности, что могут служить диагностическими элементами и маркерами их иллокутивной силы.

[10] Мне не известно, имеются ли смысловые различия между указанными вариантами жестовых реализаций. Заведомо, однако, это не омонимичные лексемы.

038-Kreidlin_G:_Layout 1 18.05.2008 15:28 Page 253

Список литературы

1. Де Йорио 1832/2000 – De Jorio, A. Gesture in Naples and Gesture in Classical Antiquity. A translation of «La mimica degli antichi investigate nel gestire napoletano» (1832), and with an Introduction and Notes, by A. Kendon. Bloomington: Indiana Univ. Press, 2000.

2. Кальбрис 1990 – Calbris, G. The semiotics of French gestures. Bloomington: Indiana Univ. Press, 1990.

3. Кендон 1995 — Kendon, A. Gestures as illocutionary and discourse structure markers in Southern Italian conversation. Journal of pragmatics, v. 23, 1995, 247 – 279.

4. Кендон 2004 – Kendon, A. Gesture: Visible Action as Utterance. Cambridge: Cambridge Univ. Press, 2004.

5. Крейдлин 2002 – Крейдлин Г. Е. Невербальная семиотика. М.: Новое литературное обозрение, 2002.\

6. Крейдлин 2005 – Крейдлин Г. Е. Мужчины и женщины в невербальной коммуникации. Мужчины и женщины в невербальной коммуникации. М.: «Языки русской культуры», 2005.

7. Крейдлин 2006 – Крейдлин Г. Е. Механизмы взаимодействия невербальных и вербальных единиц в диалоге I: Жестовые ударения // Труды международной конференции «Диалог 2006»: компьютерная лингвистика и интеллектуальные технологии». М., 2006, 290 – 296.

8. Крейдлин 2007 – Крейдлин Г. Е. Механизмы взаимодействия невербальных и вербальных единиц в диалоге: II А.Дейктические жесты и их типы // Труды международной конференции «Диалог 2006»: компьютерная лингвистика и интеллектуальные технологии». М., 2007, 300 – 327.

9. Печерская 2001 – Печерская И.Б. Семантика лекторского жеста. Дипломная работа слушателя 2 курса. Институт европейских культур РГГУ. Москва, 2001 (рукопись).

10. Поджи 1983 – Poggi, I. La mano a borsa: Analisi semantica di un gesto emblematico olofrastico // G. Attili and Pio E. Ricci-Bitti (eds.) Comunicare senza parole. Bulzoni: Roma, 1983, 219 — 238.

11. Рико-Кассар 2005 – Рико-Кассар Ф. Сопоставительный анализ невербального знакового поведения французских и русских лекторов во время лекций. Московский лингвистический журнал, т. 8, № 2, 2005, 118 – 129.

12. Эфрон 1941/1972 – Efron, D. Gesture and Environment. New York: King's Crown Press, 1941 (2nd edition – 1972: в 1972 году книга вышла под названием «Gesture, Race and Culture«).

039-Krylov_SA:_Layout 1 18.05.2008 15:28 Page 254

ИЗМЕРЕНИЕ ЧАСТОТНОСТИ СИНТАКСИЧЕСКИХ МОЛЕКУЛ (НА МАТЕРИАЛЕ ГЕНЕРАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА) [1]

EVALUATING OF FREQUENCY OF SYNTACTIC MOLECULES (ON THE EVIDENCE FROM THE RUSSIAN GENERAL CORPUS)

Крылов С.А. *(krylov-58@mail.ru)*
Институт востоковедения РАН
Институт системного анализа РАН

Сделана попытка с помощью интегрированной информационной системы StarLing подсчитать частотность синтаксических молекул (минимальных знаменательных членов предложения, способных служить ответом на вопрос) на материале Генерального корпуса русского языка (созданного на основе Уппсальского корпуса).

1. Для грамматического и лексического анализа русского языка оказывается весьма полезным понятие синтаксической молекулы (СМ) [2]. СМ есть минимальная синтаксически автономная единица членения речи, то есть минимальный отрезок, способный функционировать в качестве отдельной (быть может и эллиптичной) реплики, отвечающей на какой-либо вопрос. СМ обычно содержит не более одного полнозначного знаменательного слова; при этом в её состав может входить одно или несколько служебных (или полуслужебных) слов.

2.0. Единица, близкая синтаксической молекуле, выделяется во многих фонетических работах под названием «фонетического слова» (ФС) [3] или «морфемного комплекса» [4]. Особенности предлагаемого подхода к ФС, предполагающего составление частотного словаря фонетических слов (Крылов 2008) – такие: (а) ФС рассматривается не только в синтагматическом, но и в парадигматическом аспекте; (б) ФС трактуется как двусторонняя (знаковая) единица [5]; (в) в центре внимания находится именно инвентарный [6] (словарный, лексикологический, лексикографический) аспект ФС [7].

2.1. В защиту принимаемого подхода можно привести следующие доводы, апеллирующие к аналогии между **слитно пишущимися** («синтетическими») речевыми отрезками и **раздельно пишущимися** («аналитическими») формами:

2.1.1. Если мы признаём «**синтетические**» латинские образования с постпозитивным *–que* двусторонними единицами, то можно ли отказывать в этом статусе «**аналитическим**» русским образованиям с препозитивным *и*?

2.1.2. Если мы признаём **синтетически** выражаемые категории падежа и числа (наряду с категориями рода и одушевлённости) категориями русских прилагательных, невзирая на то, что выбор граммем этих категорий обусловлен выбором соответствующих граммем существительного, контролирующего согласование данного прилагательного (и при этом готовы исключительно на основании такого контроля усматривать наличие соответствующих граммем в означаемом словоформ несклоняемых существительных типа *кофе* или *такси*, ср.: *чёрн-**ому** кофе*, *московск-**ого** такси*), то что нам мешает распространить этот подход на категории (и граммемы), выражаемые **аналитически** (напомним, что предлоги и послелоги, в соответствии со взглядами Ю. С. Маслова, признаются аналитическими выразителями категории **падежа**)? Соответственно, логично трактовать граммему ин-эссива как часть означаемого СМ ***в** письменн-**ом*** (признаваемую у СМ ***в** стол-**е***), граммему суб-латива – как часть означаемого СМ ***под** обеденн-**ым*** (признаваемую у СМ ***под** стол-**ом***) и т.п.

[1] Работа выполнена при частичной поддержке РГНФ (проект № 07-04-00161а.

[2] См. Балли 1955, с.317.

[3] См., напр., Зиндер 1979, с. 251; Суханова 1980, с. 90; Кодзасов и Кривнова 2001, с. 27-28, 304-306. В психолингвистическом аспекте важно проводимое Е. В. Ягуновой различение между более широким классом ФС «как оперативных единиц восприятия» и более узким классом ФС как «единиц перцептивного словаря» (Ягунова 2006, с. 400).

[4] Зализняк 1985: 39.

[5] Аргументация в пользу такого подхода приведена в Крылов 2007.

[6] Об «инвентарных» единиц языка в отличие от «конструктивных» см. Касевич 1988; Крылов 2006а.

[7] Такая трактовка созвучна идеям, развиваемым санкт-петербургской школой психолингвистически адекватного моделирования речевой деятельности. С другой стороны, проблему инвентаризации составных слов ставит и успешно решает практическая лексикография (Рогожникова 1991; Рогожникова 2003).

039-Krylov_SA:_Layout 1 18.05.2008 15:28 Page 255

2.1.3. Если мы признаём двусторонними (знаковыми) единицами такие французские «**синтетические**» СМ, как *où* «, где (относит.), в котором (неодуш.)», *y* «в нём, у него (неодуш.)», *en* «от него (неодуш.)», то что нам мешает признавать двусторонними (знаковыми) единицами аналогичные им (а при известных условиях даже эквивалентные им) русские «**аналитические**» СМ: *в котором*, *у которого*, *в нём*, *у него*, *от него*, *из него* и т.п.?

2.1.4. Если мы признаём двусторонними (знаковыми) единицами такие русские «**синтетические**» образования (1а) *там*; (1б) *туда*; (1в) *оттуда*; (2а) где (относит.); (2б) *куда* (относит.); (2в) *откуда* (относит.); и т. п., то что нам мешает признать двусторонними (знаковыми) единицами их «**синтетические**» аналоги (а нередко и контекстные эквиваленты) в русском языке (1а) *в нём*, *в ней*, *на нём*, *на ней*, *у него*, *у неё*, *у них*; (1б) *в него*, *в неё*, *в них*; *на него*, *на неё*, *на них*; *к нему*, *к ней*, *к ним*; (1в) *из него*, *из неё*, *из них*; *с него*, *с неё*, *с них*; *от него*, *от неё*, *от них*; (2а) *в котором*, *в которой*, *в которых*; *на котором*, *на которой*, *на которых*; *у которого*, *у которой*, *у которых*; (2б) *в который*, *в которую*, *в которые*; *на который*, *на которую*, *на которые*; *к которому*, *к которой*, *к которым*; (2в) *из которого*, *из которой*, *с которого*, *с которой*, *с которых*; *от которого*, *от которой*, *от которых*; и т. п.)?

2.1.5. Если мы признаём двусторонней (знаковой) единицей «**синтетически**» образованную форму с деепричастным значением (ср. *читая*), то что нам мешает признать двусторонней (знаковой) единицей аналогичную ей (и нередко контекстно эквивалентную ей) «**аналитически**» образованную СМ с сочинительным значением (ср. *и читает*)?

2.2. Если мы признаём двусторонней (знаковой) единицей «**аналитически**» образованную форму с комитативным значением, оформляющую неоднородный член (напр., во фразе *Петя **с Ваней** гуляют*), то что нам мешает признать двусторонней (знаковой) единицей эквивалентную ей «**аналитически**» же образованную форму с координативным значением, оформляющую однородный член (напр., во фразе *Петя **и Ваня** гуляют*)?

3.0. СМ образуют иерархию из трёх рангов [8]: макротакты, мезотакты и микротакты.

3.1. Макротакт есть морфемный комплекс между двумя местами потенциальных пауз (в отличие от более крупной единицы - фонетической синтагмы[9], границы которой отмечены реальными паузами).

3.2. Мезотакт есть морфемный комплекс, включающий не более одного «полноударного» ФС. Мезотакт может включать в себя один или несколько **«клитикоидов»** [10] (то есть **«слабоударяемых»** ФС и «относительных клитик») – постпозитивных (**«энклитикоидов»**) или препозитивных (**«проклитикоидов»**).

3.3. Микротакт есть морфемный комплекс, содержащий ровно 1 автономный (характеризуемый единством главного словесного ударения) словесный сегмент[11]. Микротакты бывают простыми и составными. Составные микротакты включают, помимо автономного сегмента, также одну или несколько **клитик** – единиц, не несущих самостоятельного словесного ударения. Клитики подразделяются на **энклитики** (постпозитивные) и **проклитики** (препозитивные).

4.0. Инвентарь ментальных СМ (то есть синтаксических молекул, хранимых в ментальном лексиконе носителя языка) выявляется путём измерения их встречаемости в крупном корпусе текстов [12] и создания частотного инвентаря реальных СМ [13].

4.1. Эта задача может решаться по-разному [14]. Источником данных был корпус текстов, представленных в орфографической записи - Генеральный корпус русского языка (ГКРЯ), созданный на основе «Уппсальского корпуса» русского языка (УпКРЯ), составленного под руководством Л. Лённгрена (http://www.slaviska.uu.se/ryska/index.html). В 1995 гг. автором настоящей работы под руководством С. А. Старостина (1953-2005) [15] материалы УпКРЯ были преобразованы в формат текстовой базы данных, получившей название ГКРЯ [16].

5.0. В 2005-2008 гг. ГКРЯ был снабжён «грубой» разметкой тактовой делимитации. Она устроена так.

5.1. Пробелы письменного текста бывают **паузальные** (соответствующие границам макротактов в устной

[8] Эти понятия введены в Крылов 2006b, Крылов 2006c.

[9] О фонетических синтагмах см. Кодзасов и Кривнова 2001, с. 2728, 304-306.

[10] Термины из Крылов 2006b, Крылов 2006c.

[11] Иными словами, микротакт соответствует «тактовой группе» (Чурганова 1973, с. 22-23, 27, 31-32), «акцентному слову» (Маслов 1975, с. 91-93), «акцентной группе» (Зиндер 1979, с. 251) «речевому такту» (Кодзасов и Кривнова 2001, с. 306), «ритмической группе» (Зиндер 1979, с. 250, 262; Кодзасов и Кривнова 2001, с. 310).

[12] Т.о., разрабатывается «структурно-вероятностная модель» языка (Шайкевич 1990: 231).

[13] См. Крылов и Ягунова 2006.

[14] В современной корпусной лингвистике иногда не ограничиваются членением текста на графические слова (от пробела до пробела), а выделяют также более сложные единицы – т.н. «составные слова». См., напр., Венцов и др. 2004a; Венцов и др. 2004b; Копотев 2004; Мустайоки и Копотев 2004; Ягунова 2006.

[15] Взгляды С. А. Старостина на задачи корпусной филологии отразились в публикации Перцов и Старостин 1995; см. также проект ПРИОР «Портал “Русский язык и литература”» (http://prior.russia-gateway.ru/content/view/1023/150/).

[16] См. Крылов и Старостин 2005.

039-Krylov_SA:_Layout 1 18.05.2008 15:28 Page 256

речи) и **беспаузальные** (для транскрибирования которых использован создан набор из 6 искусственных делимитаторов [17]:

{ после проклитик;

} перед энклитикой;

< после проклитикоида;

> перед энклитикоидом;

<> между частями мезотакта с «неустойчивым» центром (то есть сочетания, допускающего двоякую акцентуацию: либо как «клитикоид + полноударное», либо как «полноударное + клитикоид»);

+ между мезотактами, образующими один макротакт.

6. Частотность СМ в русском языке.

В результате разметки ГКРЯ оказалось возможным извлечь из него сведения о частотах СМ.

Сосредоточим внимание на одном из классов СМ – а именно, на СМ, начинающихся с проклитики. Рассмотрим небольшую «верхушку» из частотного словаря (ЧС) таких СМ.

В таблице столбец (А) указывает на инвентаризуемую СМ (макротакт), (Б) - на её относительную частотность по числу текстов (%), (В) - на её абсолютную частотность по числу текстов, (Г) - на её ранг в ЧС, упорядоченном по числу текстов (этот параметр в таблице является ключевым), (Д) - на её относительную частотность по числу вхождений при измерении общего числа вхождений СМ в корпус (в числе вхождений данной единицы на 10 тыс. [18]), (Е) - на её абсолютную частотность по числу вхождений (этот параметр в таблице является побочным), (Ж) - на её ранг в ЧС, упорядоченном по числу вхождений. Под ключевым параметром понимается та из числовых характеристик СМ, которая лежит в основе упорядочения строк в таблице (в нашем случае в качестве ключевого был выбран параметр Г).

Для наглядности ниже дана лишь частотная «верхушка» одного из полученных словарей [19].

Частотность мезотактов с проклитиками в ЧС макротактов.

А	Б	В	Г	Д	Е	Ж
о{том	35.98	204	61	44.75	319	86
у{нас	30.51	173	87	44.61	318	87
из{них	25.93	147	121	28.90	206	161
об{этом	25.75	146	124	28.90	206	163
не{}было	24.34	138	138	48.26	344	76
в{нем	22.22	126	165	28.20	201	170
и{все	21.87	124	168	29.60	211	155
и{это	20.99	119	190	22.31	159	249
у{него	20.28	115	200	43.35	309	90
а{потом	19.93	113	207	30.72	219	145
и{другие	19.75	112	219	18.80	134	316
с{ним	19.58	111	220	26.37	188	186
к{нему	19.05	108	231	24.97	178	210
в{ней	18.87	107	235	20.06	143	286
и{его	18.87	107	236	17.11	122	378
в{котором	18.17	103	247	17.82	127	352
у{них	17.81	101	254	21.60	154	263
в{частности	17.46	99	264	22.87	163	241
и{что	16.93	96	286	19.36	138	302
к{сожалению	16.75	95	301	16.41	117	404
на{него	16.05	91	312	22.59	161	243
у{нее	15.87	90	319	29.60	211	157
у{меня	15.87	90	320	25.11	179	208
и{как	15.52	88	332	16.27	116	408
до{сих{пор	15.52	88	336	15.43	110	451
к{ней	15.34	87	340	19.64	140	292
и{других	15.34	87	344	15.43	110	453
не{может	15.34	87	346	14.73	105	480
в{них	15.17	86	349	14.87	106	468
в{целом	14.81	84	359	15.85	113	428

[17] О просодических швах разной глубины см. Кривнова 2007: 60-69.

[18] Ср. понятие «ipm» (“instances per million words”) в Sharoff 2002 (http://www.artint.ru/projects/frqlist.asp).

[19] Развёрнутые версии этой и других таблиц (проекция срезов ЧС 4096 наиболее частых макротактов и 4096 самых частых микротактов) выложены в Сети (http://www.yazykoznanie.narod.ru/PHOWO08.html).

039-Krylov_SA:_Layout 1 18.05.2008 15:28 Page 257

на{себя	14.81	84	360	15.57	111	443
на{них	14.64	83	368	14.59	104	486
к{тому}же	14.64	83	369	13.19	94	550
а{это	14.46	82	375	13.61	97	530
и{так	14.11	80	390	15.71	112	437
в{мире	14.11	80	393	14.17	101	501
а{что	13.76	78	405	14.45	103	490
в{*Москве	13.58	77	413	13.61	97	531
и{вдруг	13.23	75	427	17.40	124	370
в{стране	13.23	75	428	16.69	119	393
в{год	13.23	75	430	15.43	110	446
в{которой	13.23	75	440	12.06	86	621
к{ним	12.87	73	456	12.49	89	591
в{сторону	12.70	72	465	13.75	98	521
и{снова	12.52	71	471	14.73	105	479
и{тогда	12.35	70	485	14.03	100	506
и{они	12.35	70	486	13.75	98	522
во{всех	12.35	70	489	10.80	77	726
и{т#+д#	12.17	69	491	15.43	110	454
а{он	12.17	69	493	14.59	104	483
в{жизни	12.17	69	496	12.91	92	564
как{правило	12.17	69	498	12.20	87	610
не{будет	12.17	69	501	11.36	81	678
не{мог	11.82	67	524	15.15	108	462
и{теперь	11.82	67	526	12.49	89	590
в{которых	11.82	67	531	10.80	77	724
и{она	11.64	66	533	17.25	123	373
а{затем	11.46	65	564	10.94	78	710
от{него	11.29	64	570	13.19	94	554
к{себе	11.29	64	574	11.50	82	663
в{результате	11.29	64	576	10.52	75	752
с{ними	11.11	63	585	11.78	84	647
к{примеру	11.11	63	590	11.08	79	697
во{всем	11.11	63	592	10.66	76	737
а{я	10.93	62	597	14.73	105	475
в{себе	10.93	62	605	11.36	81	673
в{первую+очередь	10.93	62	612	9.96	71	821
и{потому	10.76	61	620	10.52	75	756
а{теперь	10.76	61	621	10.38	74	769
в{основном	10.76	61	625	9.68	69	857
и{тут	10.58	60	631	11.50	82	662
и{их	10.58	60	635	9.82	70	842
и{когда	10.41	59	651	11.64	83	652
с{ней	10.41	59	652	11.50	82	668
в{чем	10.41	59	657	10.24	73	787
для{него	10.41	59	661	9.96	71	825
на{нее	10.23	58	671	12.63	90	586
и{ее	10.23	58	679	10.24	73	788
а{когда	10.05	57	699	10.52	75	750
и{сейчас	10.05	57	704	8.70	62	1001
и{я	9.88	56	706	23.15	165	233
о{чем	9.70	55	741	10.66	76	745
в{нашей<стране	9.52	54	762	9.40	67	896
до{конца	9.52	54	765	9.26	66	914
по{существу	9.52	54	768	8.42	60	1049
а{тут	9.35	53	779	9.12	65	932
не{так	9.35	53	781	8.70	62	1006
для{себя	9.17	52	807	8.84	63	978
в{прошлом<>году	8.99	51	829	9.54	68	880
от{нее	8.99	51	833	8.84	63	987
не{знаю	8.82	50	850	9.96	71	832
а{она	8.82	50	852	9.40	67	894
не{раз	8.82	50	860	8.56	61	1026
на{месте	8.82	50	863	8.28	59	1070
тем<не{менее	8.82	50	865	8.28	59	1081
во{многом	8.82	50	866	8.14	58	1087
не{могут	8.82	50	870	8.00	57	1119
и{уже	8.82	50	872	7.86	56	1149
от{них	8.82	50	874	7.58	54	1230
в{последнее<>время	8.64	49	902	7.86	56	1142
из{которых	8.64	49	905	7.72	55	1177
в{*С*С*С*Р	8.47	48	913	10.66	76	736
в{общем	8.47	48	917	9.26	66	911

039-Krylov_SA:_Layout 1 18.05.2008 15:28 Page 258

в{руках	8.47	48	923	8.28	59	1060
а{значит	8.47	48	931	7.44	53	1239
для{них	8.47	48	932	7.29	52	1288
а{может	8.29	47	941	8.84	63	974
а{ты	7.94	45	983	10.94	78	711
в{конце<>концов	7.94	45	994	8.28	59	1059
и{все}же	7.94	45	995	8.28	59	1062
и{мы	7.94	45	996	8.28	59	1063
с{собой	7.94	45	997	8.28	59	1077
и{вообще	7.94	45	1001	7.72	55	1174
и{сам	7.94	45	1002	7.72	55	1175
для{всех	7.94	45	1012	7.15	51	1324
и{наконец	7.94	45	1013	7.15	51	1326
не{надо	7.76	44	1036	8.00	57	1120
на{землю	7.76	44	1039	7.72	55	1183
в{одном	7.76	44	1050	7.15	51	1319
в{самом	7.76	44	1052	7.01	50	1351
в{то}же<>время	7.76	44	1062	6.59	47	1464
не{всегда	7.76	44	1063	6.45	46	1527
в{работе	7.58	43	1077	7.72	55	1167
о{нем	7.58	43	1080	7.44	53	1261
и{тут}же	7.58	43	1082	7.29	52	1290
на{все	7.58	43	1086	6.87	49	1406
в{свое<время	7.58	43	1088	6.59	47	1463
в{*С*Ш*А	7.41	42	1103	8.00	57	1110
а{как	7.41	42	1113	7.15	51	1316
во{все	7.41	42	1126	6.45	46	1505
на{котором	7.41	42	1127	6.45	46	1524
в{таких	7.41	42	1128	6.31	45	1549
на{нем	7.41	42	1132	6.17	44	1600
о{них	7.41	42	1133	6.17	44	1606
за{ним	7.23	41	1141	9.12	65	935
и{все-таки	7.23	41	1154	7.01	50	1356
на{другой	7.23	41	1157	6.87	49	1407
на{этом	7.23	41	1158	6.73	48	1442
во{всяком<>случае	7.23	41	1162	6.45	46	1506
на{меня	7.05	40	1175	8.98	64	960
у{вас	7.05	40	1180	8.00	57	1133
на{улице	7.05	40	1185	7.58	54	1221
со{мной	7.05	40	1188	7.44	53	1269
о{себе	7.05	40	1191	7.29	52	1293
в{глаза	7.05	40	1195	7.15	51	1318
к{чему	7.05	40	1214	6.03	43	1654
в{городе	6.88	39	1219	8.98	64	954
для{меня	6.88	39	1223	8.28	59	1061
и{т#п#	6.88	39	1224	8.28	59	1064
и{стал	6.88	39	1229	7.72	55	1176
в{настоящее<>время	6.88	39	1234	7.29	52	1280
не{знал	6.88	39	1237	7.01	50	1365
на{берегу	6.88	39	1239	6.87	49	1405
и{там	6.88	39	1244	6.59	47	1469
а{сейчас	6.88	39	1249	6.45	46	1500
в{него	6.88	39	1252	6.31	45	1548
ко{мне	6.70	38	1270	8.00	57	1118
на{земле	6.70	38	1279	7.01	50	1363
на{работу	6.70	38	1280	7.01	50	1364
в{день	6.70	38	1286	6.73	48	1427
для{нас	6.70	38	1289	6.59	47	1467
на{всех	6.70	38	1301	6.03	43	1658
не{все	6.70	38	1302	6.03	43	1662
не{хватает	6.70	38	1306	5.89	42	1710
в{школе	6.53	37	1315	11.08	79	694
с{тех<>пор	6.53	37	1335	6.59	47	1491
с{одной>стороны	6.53	37	1341	6.31	45	1567
и{в{то}же<>время	6.53	37	1344	6.17	44	1593
из{нас	6.53	37	1345	6.17	44	1595
со{всеми	6.53	37	1346	6.17	44	1627
на{этот<раз	6.35	36	1386	6.59	47	1475
а{может<>быть	6.35	36	1400	5.89	42	1682
в{себя	6.35	36	1405	5.75	41	1742
и{есть	6.35	36	1407	5.75	41	1759
за{собой	6.35	36	1414	5.61	40	1816
в{нее	6.35	36	1416	5.47	39	1865

039-Krylov_SA:_Layout 1 18.05.2008 15:28 Page 259

и{поэтому	6.35	36	1418	5.47	39	1880
из{этих	6.35	36	1419	5.47	39	1881
а{потому	6.17	35	1437	7.58	54	1199
в{воздухе	6.17	35	1446	6.59	47	1462
в{памяти	6.17	35	1454	6.03	43	1637
и{тоже	6.17	35	1458	5.89	42	1697
в{разных	6.17	35	1462	5.75	41	1741
не{может<>быть	6.17	35	1463	5.75	41	1770
в{этом<случае	6.17	35	1477	5.33	38	1929
о{котором	6.17	35	1479	5.33	38	1959
в{свою<очередь	6.17	35	1481	5.19	37	2009
с{таким	6.17	35	1484	5.05	36	2130
у{тебя	6.00	34	1494	7.01	50	1380
к{нам	6.00	34	1498	6.45	46	1516
не{могу	6.00	34	1503	6.17	44	1603
и{сразу	6.00	34	1519	5.75	41	1760
с{которой	6.00	34	1524	5.47	39	1902
в{данном<>случае	6.00	34	1525	5.33	38	1925
по{всем	6.00	34	1529	5.33	38	1965
на{всю	6.00	34	1536	5.05	36	2096
со{временем	6.00	34	1540	5.05	36	2134
то<и{дело	5.82	33	1561	5.89	42	1728
на{которой	5.82	33	1568	5.61	40	1832
в{самом<>деле	5.82	33	1574	5.47	39	1866
друг<от{друга	5.82	33	1576	5.47	39	1876
не{случайно	5.82	33	1577	5.47	39	1888
в{последние<>годы	5.82	33	1579	5.33	38	1927
на{ней	5.82	33	1581	5.33	38	1953
в{два	5.82	33	1587	5.19	37	2005
у{всех	5.82	33	1592	5.19	37	2053
а{они	5.82	33	1594	5.05	36	2062
на{что	5.82	33	1597	5.05	36	2097
с{другой<>стороны	5.82	33	1604	4.91	35	2212
в{доме	5.64	32	1614	7.86	56	1141
и{опять	5.64	32	1617	7.44	53	1250
на{свете	5.64	32	1623	6.45	46	1525
не{могла	5.64	32	1627	6.17	44	1602
на{самом<>деле	5.64	32	1633	5.89	42	1708
в{лицо	5.64	32	1637	5.61	40	1809
в{сущности	5.64	32	1640	5.47	39	1867
в{одну	5.64	32	1653	5.19	37	2007
ни{разу	5.64	32	1656	5.19	37	2027
в{дальнейшем	5.64	32	1662	5.05	36	2063
в{одной	5.64	32	1663	5.05	36	2065
в{руки	5.64	32	1664	5.05	36	2066
и{потом	5.64	32	1673	4.91	35	2179
не{имеет	5.64	32	1675	4.91	35	2191
и{больше	5.64	32	1680	4.77	34	2250
а{где	5.64	32	1687	4.63	33	2304
не{могли	5.64	32	1688	4.63	33	2342
в{комнате	5.47	31	1711	5.89	42	1686
друг<с{другом	5.47	31	1712	5.89	42	1694
в{истории	5.47	31	1719	5.75	41	1739
и{сказал	5.47	31	1724	5.61	40	1821
а{мы	5.47	31	1730	5.33	38	1922
в{другую	5.47	31	1740	5.05	36	2064
по{крайней<>мере	5.47	31	1745	5.05	36	2115
и{дальше	5.47	31	1763	4.63	33	2326
с{трудом	5.47	31	1767	4.63	33	2363
на{улицу	5.29	30	1775	6.87	49	1408
а{вы	5.29	30	1786	6.03	43	1635
в{природе	5.29	30	1793	5.75	41	1740
на{столе	5.29	30	1801	5.61	40	1833
а{главное	5.29	30	1813	5.05	36	2061
о{своем	5.29	30	1814	5.05	36	2101
в{среднем	5.29	30	1820	4.91	35	2161
а{сам	5.29	30	1823	4.77	34	2231
на{одном	5.29	30	1826	4.77	34	2253
и{тем<не{менее	5.29	30	1831	4.63	33	2327
в{голову	5.29	30	1835	4.49	32	2394
в{различных	5.29	30	1836	4.49	32	2396
в{системе	5.29	30	1837	4.49	32	2397
время<от{времени	5.11	29	1868	5.47	39	1870

039-Krylov_SA:_Layout 1 18.05.2008 15:28 Page 260

в{*Москву	5.11	29	1871	5.33	38	1924
в{один	5.11	29	1893	4.91	35	2160
с{детства	5.11	29	1897	4.91	35	2211
а{все	5.11	29	1917	4.49	32	2387
в{этом<году	5.11	29	1920	4.49	32	2399
в{другом	5.11	29	1932	4.35	31	2489
а{иногда	5.11	29	1937	4.21	30	2564
в{каждой	5.11	29	1938	4.21	30	2570
в{этом	5.11	29	1939	4.21	30	2573
к{концу	5.11	29	1943	4.21	30	2607
не{столько	5.11	29	1945	4.21	30	2626

Заключение

Приведённые данные являются сугубо предварительными, так как процесс разметки ГКРЯ пока продолжается. Многие из выделенных синтаксических молекул нуждаются в более тщательной интерпретации. Но можно надеяться, что и эта предварительная стадия разметки способна привести к получению ценной информации о статистике употребления СМ в русском языке.

Список литературы

1. Аванесов Р. И. Русское литературное произношение. М.: Просвещение, 1968.- 288 с.
2. Аванесов Р. И. Русская литературная и диалектная фонетика. М.: Просвещение, 1974.- 287 с.
3. Балли Ш. Общая лингвистика и вопросы французского языка. М.: ИЛ, 1955.- 416 с.
4. Венцов А. В., Касевич В. Б., Ягунова Е. В. Идиома, слово, фонетическое слово // Язык и речь: проблемы и решения. Сб. научных трудов к юбилею проф. Л. В. Златоустовой. М.: Изд-во МГУ, 2004a. С. 357-363.
5. Венцов А. В., Грудева Е. В., Касевич В. Б., Ягунова Е. В. Об идиомах в национальном корпусе русского литературного языка // Корпусная лингвистика-2004. Тезисы международной конференции. 12-14 октября 2004 г., СПб.: Изд. СпбГУ, 2004b. С. 1718.
6. Высотский С. С. Звук речи в контексте // Диалектологические исследования по русскому языку. М.: Наука, 1977. С. 2438.
7. Зализняк А. А. От праславянской акцентуации к русской. М.: Наука, 1985.- 428 с.
8. Зиндер Л. Р. Общая фонетика. Изд. 2-е. М.: Высшая школа, 1979.- 312 с.
9. Касаткин Л. Л. Фонетика современного русского литературного языка. М.: Изд-во МГУ, 2003.- 223 с.
10. Касевич В. Б. Семантика. Синтаксис. Морфология. М.: Восточная литература, 1988.
11. Кодзасов С. В., Кривнова О. Ф. Общая фонетика. М.: РГГУ, 2001.- 592 с.
12. Копотев М. Несмотря на, потому что, или многокомпонентные единицы в аннотированном корпусе русских текстов // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог'2004» («Верхневолжский», 2-7 июня 2004 г.), М., Наука, 2004.
13. Кривнова О. Ф.Ритмизация и интонационное членение текста в «процессе речи-мысли». ДД. М., 2007.
14. Крылов С. А.Об инвентарных и конструктивных единицах языка // Язык и речевая деятельность. 2003. Вып. 6. СПб.: Филол. ф-т СПбГУ, 2006a. С. 926.
15. Крылов С. А.Фонетическое слово и его корреляты в русском письменном тексте (с точки зрения корпусной лингвистики) // Корпусная лингвистика-2006. Труды международной конференции. 10-14 октября 2006 г., СПб.: Изд. СпбГУ РХГА, 2006b. С. 190200.
16. Крылов С. А. Фонетическое слово и его корреляты в русском письменном тексте (с точки зрения корпусной лингвистики) // Пятая выездная школа-семинар «Порождение и восприятие речи». Череповец: ЧГУ – ПЛО, 2006c. С. 6696.
17. Крылов С. А. Фонетическое слово в семиотическом аспекте // Фонетика сегодня. Материалы докладов и сообщений V международной научной конференции. 8-10 октября. М.: ИРЯ РАН, 2007. С. 210212.
18. Крылов С. А., Старостин С. А. Металингвистическая разметка текстовых баз данных в системе STARLING и современные задачи корпусной лингвистики // Прикладная лингвистика в поиске новых путей. Тезисы Международной научной конференции MegaLing'2005. Крым, Украина. 27 июня - 2 июля 2005 г. Симферополь: ТЭИ, 2005.
19. Крылов С. А. , Ягунова Е. В. 2006. Квантитативный подход к выделению инвентарных единиц языка // Материалы 2-й международной конференции по когнитивной науке. СПб., 9-13 июня 2006 года. СПб., 2006. С. 329-330.
20. Маслов Ю. С. Введение в языкознание. М.: Высшая школа, 1975.

039-Krylov_SA:_Layout 1 18.05.2008 15:28 Page 261

21. Мустайоки А., Копотев М. К вопросу о статусе эквивалентов слова типа потому что, в зависимости от, к сожалению // Вопросы языкознания, 2004, № 3.

22. Перцов Н.В., Старостин С. А. О лексикографической справочной информационной системе ЛЕКСИС по русскому языку // Труды Международного семинара «Диалог'95» по компьютерной лингвистике и ее приложениям = «Dialogue'95. Computational linguistics and its applications» international workshop, Казань, 31 мая - 4 июня 1995 г. - Казань, 1995. - С. 247.

23. Рогожникова Р. П. Словарь эквивалентов слова: наречные, служебные, модальные единства. - М.: Русский язык, 1991. - 255 с.

24. Рогожникова Р.П. Толковый словарь словосочетаний, эквивалентных слову. М., 2003.

25. Русская грамматика. Т. 1. М.: Наука, 1980.- 783 с.

26. Суханова М. С. Основные сведения об ударении // Русская грамматика 1980. С. 9095.

27. Чурганова В. Г. Очерк русской морфонологии. М.: Наука, 1973.- 239 с.

28. Шайкевич А. Я. Количественные методы в языкознании // Лингвистический энциклопедический словарь. М.: Советская энциклопедия, 1990. С. 231.

29. Ягунова Е. В. Неоднословные целостности в словаре и в корпусе // Корпусная лингвистика-2006. Труды международной конференции. 10-14 октября 2006 г., СПб.: Изд. СпбГУ РХГА, 2006. С. 395412.

30. Sharoff, Serge, Meaning as use: exploitation of aligned corpora for the contrastive study of lexical semantics. // Proc. of Language Resources and Evaluation Conference (LREC02). May, 2002, Las Palmas, Spain, 2002.

040-Krylova_T:_Layout 1 18.05.2008 15:30 Page 262

БЛАГОРОДНЫЙ: НАИВНО-ЯЗЫКОВЫЕ ПРЕДСТАВЛЕНИЯ О СВЯЗИ МЕЖДУ ВНУТРЕННИМИ КАЧЕСТВАМИ И СОЦИАЛЬНЫМ ПРОИСХОЖДЕНИЕМ ЧЕЛОВЕКА[1]

БЛАГОРОДНЫЙ: LANGUAGE CONCEPTION OF CONNECTION BETWEEN INTERNAL QUALITIES AND BIRTH OF PERSON

Крылова Т.В. *(ta-kr@yandex.ru)*
Институт русского языка им. В.В.Виноградова РАН

В статье рассматриваются прилагательные *благородный* и *великодушный* в их основном, «этическом» значении (*благородный 1.1* и *великодушный 1*), описываются различия между ними. В частности, обнаруживается, что *великодушный*, в отличие от *благородный*, указывает на снисхождение субъекта к партнеру по общению. Делается попытка установить связь между семантикой лексемы *благородный 1.1* и его внутренней формой (так, *благородный человек*, в отличие от *великодушный человек*, указывает на врожденное свойство). Дается описание многозначности прилагательного *благородный*. На основании анализа ряда значений этого слова - *благородный 3.1* (*благородное лицо*) и *благородный 3.2* (*благородное животное*) – высказывается предположение о том, что в современном языковом сознании сохраняются представления о связи между внутренними качествами и социальным происхождением человека.

В данной статье используется метод сопоставительного семантического анализа близких лексем, применяемый в НОСС, с привлечением в качестве материала данных сочетаемости и синтаксиса. Материалы статьи могут быть полезны как для словарей синонимов, таких, как НОСС, так и для толковых словарей – вследствие наличия в ней описания многозначности изучаемых слов.

Рассмотрим слова *великодушный* и *благородный*[2]. И те, и другие в своем основном значении (*благородный 1.1* и *великодушный 1)* обозначают качество человека, способного поступать вопреки своим личным интересам. Однако семантика и внутренняя форма этих слов существенно различаются.

Великодушный может толковаться как ‘такой человек, который из-за своей доброты делает другому человеку что-л. хорошее, хотя ему самому это не приносит выгоды и, напротив, может причинить неудобства, или же, вопреки ожиданиям, не использует свое право или возможность сделать ему что-л. плохое’.

Во внутренней форме слова *великодушный* используется метафора большого размера души, которая в наивно-языковом сознании связывается с представлением о доброте – достаточно вспомнить хотя бы *человека большой души.* (При этом интересно, что «ширина» души связывается с одним определенным аспектом доброты, а именно, с щедростью, готовностью отдавать; ср. *широкая натура <душа>*, *широкий жест; Хоть Шолохов был грубоват, но вместе с тем отзывчив, щедр, как говорится, человек* «*широкой души*«. *Когда он отмечал свой очередной юбилей в ресторане ростовской гостиницы, то пригласил на банкет всех ее жильцов* («Лебедь» (Бостон), 2003.10.19); *С 1 декабря зарплаты учителей и медиков увеличиваются в два раза, остальных работников бюджетной сферы - в среднем в 1,4. Регионы, которые несут на себе основную социальную нагрузку, оплатить широкий жест федеральных властей не в состоянии* («Известия», 2001.11.30). Ср. также стандартные сочетания *добрый и великодушный, щедрый и великодушный*.

Великодушный может указывать на два аспекта доброты – щедрость и милосердие. Этому соответствуют два режима употребления этого слова (и его дериватов).

[1] Данная работа осуществлена при поддержке гранта РГНФ №06-04-00289а «Разработка словника и проспекта Активного словаря русского языка» на 2006–2008 гг., гранта Программы фундаментальных исследований ОИФН РАН «Русская культура в мировой истории», гранта Президента РФ для поддержки научных исследований, проводимых ведущими научными школами РФ № НШ-5611.2006.6.

[2] Хотя речь будет идти прежде всего о прилагательных *великодушный* и *благородный*, мы будем привлекать для рассмотрения также их адвербиальные и субстантивные дериваты. Кроме того, хотя объектом непосредственного рассмотрения являются лексемы *великодушный 1* и *благородный 1.1*, мы будем иногда использовать также примеры на лексемы *великодушный 2* (*великодушное предложение*) и *благородный 1.2* (*благородное дело*), очень близкие к ним по значению.

040-Krylova_T:_Layout 1 18.05.2008 15:30 Page 263

1) Субъект, повинуясь движению души, делает другому человеку что-л. хорошее, несмотря на то, что это идет вразрез с его интересами. В этом случае можно говорить о своеобразной щедрости: субъекту «не жалко» что-л. сделать для другого, хотя это не приносит ему самому выгоды, или даже создает для него неудобства; ср. *великодушно предложил <уступил, разрешил, приютил, пригласил, пообещал, похвалил>*; ср. также выражение *борьба великодуший.* Ср. *Великодушно предложил помочь; Когда директор подошел ко мне и спросил, не остались ли у меня места, я великодушно уступил ему восемь контрамарок* (Г. Васильев, Роли, которые нас выбирают); *Несколько раз его тоже ночью выводили «на волю» и, насладившись изуверскими шутками, великодушно отпускали дрыхнуть* («Жизнь национальностей», 2002).

При этом, в соответствии с идеей щедрости, центральным для этого режима употребления является смысл ‘дать’: чаще всего *великодушный* человек дает партнеру по общению какой-л. объект или предоставляет ему возможность сделать что-л.; ср. *Каждый месяц он высылал ей по 75 рублей, а когда она написала ему, что задолжала художникам 100 рублей, то он прислал ей и эти сто. Какой добрый, великодушный человек!* (А. П. Чехов, Попрыгунья); *Сегодня ее день, соседка, та, что торгует картофелем, тряпочкой вытирает кровь с ее плеча, а продавец колбас великодушно протягивает сверток, упоительно пахнущий свежей, розовой бужениной* (У. Нова, Инка); *Вице-спикер Государственной думы Александр Венгеровский неоднократно содействовал новоявленному государю, за что тот великодушно даровал ему княжеский титул* («Вслух о.», 2003); см. также ↑.

2) Субъект, движимый жалостью, не совершает ожидаемых плохих действий в отношении другого человека[3], которые являются вполне закономерными, учитывая: а) вину этого человека перед субъектом; б) антагонистические отношения между ними (враги, соперники, противники и т. д.). В первом случае субъект не использует свое моральное право на нанесение ущерба, т. е. воздерживается от наказания, во втором случае - не использует свою возможность нанести ущерб.

Особенно характерны для *великодушный* и его дериватов случаи первого типа; ср. *Рабы оказались великодушнее охраны. Они не убили их, не избили, они велели им только раздеться, разуться и босиком в нижнее белье отпустили* (А, Солженицын, Архипелаг ГУЛАГ); *Амнистии! Великодушной и широкой амнистии жаждали мы!* (А, Солженицын, Архипелаг ГУЛАГ); *Потом в «прикормленной» Кремлем газете подробно и подчеркнуто зло описывалось, – как был прощен генерал этим самым великодушным министром и допущен в то место, где пьют «рюмку чая»* («Советская Россия», 2003.08.09). При этом *великодушный* используется не только тогда, когда субъект воздерживается от наказания, но и тогда, когда он не испытывает по отношению к провинившемуся плохих чувств – не сердится, не обижается и пр.; ср. *А если новая дружба или любовь окажется сплошным разочарованием, будьте великодушны и простите людей, оказавшихся недостойными вашего чувства* («100% здоровья», 2003); - *Всякий может ошибиться, - великодушно заметил Волька, с сочувствием глядя на сконфуженного Хоттабыча* (Л. Лагин, Старик Хоттабыч); *Я тоже ее люблю, а потому через некоторое время я зову ее, она возвращается и великодушно прощает меня* (Г. Вишневская, Галина. История жизни); ср. также уходящую фразему *простите великодушно*.

Случаи второго типа представлены в контекстах типа *Разбойники проявили великодушие и отпустили его; Боксер проявил великодушие и пощадил соперника.*

Интересная особенность слова *великодушный* и его дериватов состоит в том, что они чаще всего предполагают преимущество субъекта перед партнером по общению, по отношению к которому он ведет себя великодушно. Природа этого преимущества может быть различной. Перечислим некоторые, наиболее типичные случаи.

а) Субъект сильнее партнера по общению или имеет более высокий статус; ср. *В честь пра́здника па́схи одному́ из осуждённых, по вы́бору Ма́лого Синедрио́на и по утвержде́нию ри́мской вла́сти, великоду́шный ке́сарь импера́тор возвраща́ет его презре́нную жизнь!* (М. Булгаков, Мастер и Маргарита); *В кружке Айши бренчали уже четыре лиры, одиннадцать сольди и кольцо из американского золота, великодушно снятое с пальца некоей маркизой Нукапрути* (И. Эренбург, Необычайные похождения Хулио Хуренито).

б) Субъект играет более «выигрышную» роль в ситуации общения (роль хозяина в ситуации приема гостей, роль адресата просьбы в ситуации просьбы и пр.); ср. *– Можно сигарету? – Бери две, - великодушно разрешил он; Тем не менее гроссмейстер великодушно позволил корреспондентам журнала «Нате вам» Гайкину и Шурупову вторгнуться в его жизнь и святая святых - творческую лабораторию* («Шахматное обозрение», 2004); *великодушно приютить <пригласить>*.

в) Субъект имеет моральное преимущество перед партнером по общению. *Вы смеле́е, честне́е, глу́бже нас, но вду́майтесь, бу́дьте великоду́шны хоть на ко́нчике па́льца, пощади́те меня́* (А. П. Чехов, Вишневый сад). Сюда относятся также случаи типа *великодушно простить,* когда моральное преимущество субъекта основано на том, что его партнер по общению виноват перед ним; - *Извините, - сказал он угрюмо. – Тут у нас вчера немного… - Ну, дело житейское, - великодушно махнул рукой Нейман* (Ю. Домбровский, Факультет ненужных вещей).

[3] Оба этих режима употреблений смыкаются в случае *великодушно даровать прощение*: в данном котнексте в значении *великодушно* одновременно реализуется и идея милосердия, и идея щедрости.

040-Krylova_T:_Layout 1 18.05.2008 15:30 Page 264

г) Субъект в чем-то одержал победу над партнером по общению; ср. *Победы, одержанные им над Экселенцем, были настолько велики и очевидны, что он, без сомнения, мог позволить себе быть великодушным* (А. и Б. Стругацкие, Жук в муравейнике); *Мустафа, - сказал он наконец, обернувшись к другу, - теперь я признаю, что ты был лучшим лошадником, но ты знаешь, что и я любил лошадей и кое-что в них понимал? – А кто этого не знает? – великодушно воскликнул Мустафа* (Ф. Искандер, Сандро из Чегема).

Как бы то ни было, субъект в случае *великодушный* в чем-то превосходит партнера по общению, так что его добрый поступок расценивается как снисхождение сильного к слабому; уходящая фразема *простите великодушно* как раз и обозначает просьбу о снисхождении. Ср. также *Она писала: «Нация, удочерившая меня, ты великодушна и бесконечно снисходительна к иностранцам, но в глубине души всегда бываешь несколько удивлена, когда обнаруживаешь, что иноплеменники пользуются ножами и вилками так же как и ты»* («Вестник США», 2003.06.25).

Благородный скорее следует толковать как ‘такой, который руководствуется в своих поступках не своими личными интересами, а интересами других людей или отвлеченными представлениями о том, что хорошо, а что плохо’.

Это слово (и его дериваты) описывает два типа ситуаций:

1) Субъект делает что-л. хорошее, хотя это не приносит ему выгоды или даже приносит вред; *Где́-то у Вундта, что ли, я чита́л, что когда́ по́лчища муравьёв встреча́ют по доро́ге руче́й, то пе́рвые ряды́ броса́ются в него́ и застила́ют свои́ми тела́ми, а остальные прохо́дят по их тру́пам и иду́т да́льше. Благоро́дно? О́чень благоро́дно*, *коне́чно* (Ю. Домбровский, Обезьяна приходит за своим черепом); *Какой еще политес, когда десять лет назад в холодном и мрачном коричневом коридоре она засовывала себе в лифчик контрольную, которую он благородно вынес из триста четырнадцатой аудитории, где всегда проходили экзамены по физике!* (Т. Устинова, Подруга особого назначения); *Старшего стыдят и обличают, с ним не ленятся серьезно говорить и объяснят, почему это нужно - защищать и оберегать слабого, и как это благородно* («Семейный доктор»).

2) Субъект не делает чего-л. плохого (или нежелательного для кого-л.), хотя по какой-то причине это может быть желательно для него самого (выгодно, интересно и пр.); ср. *В су́щности, ушёл да́же дово́льно благоро́дно: разгово́ров о разме́не кварти́ры не затева́л, а ведь мог бы...* (И. Грекова, Перелом); *Говоря́т, что в своё вре́мя Резерфорд дал кля́тву не начина́ть рабо́тать с нейтро́нами, опаса́ясь, что с их по́мощью мо́жно бу́дет добра́ться до огро́мных взрывны́х сил. Благоро́дно, но чистоплю́йство бессмы́сленное* (В. Гроссман, Жизнь и судьба); *Узнав, что Анна нравится его брату, Иван, как человек благородный, прекратил свои ухаживания.*

Различия между рассматриваемыми словами можно вкратце описать так.

1) В случае *великодушный* всегда присутствует конкретный партнер по общению, которому «адресованы» поступки субъекта. Для *благородный* это не обязательно; *благородный* человек может действовать не в интересах другого человека, а во имя принципов. Ср. *Впро́чем, и «осьмино́гов» - так называ́лись благоро́дные и чи́стые ры́цари нау́ки, кого́ интересова́ло всё равно́ что, но непреме́нно выходя́щее за ра́мки шко́льной прему́дрости - бы́ло не так уж мно́го* (Ю. Трифонов, Дом на набережной); *Ведь э́то всё равно́ что назва́ть гла́вным геро́ем рома́нов Георгия Маркова благоро́дного диссиде́нта-правозащитника, бо́рющегося с коммунисти́ческим режи́мом* (Дискуссия о научной фантастике). Ср. также разнообразные употребления лексемы *благородный 1.2,* такие, как *благородная миссия, благородная тяга к знаниям, благородное дело просвещения,* а также устойчивое выражение *благородное негодование* [в этом случае подразумевается, что субъекта возмущает какой-л. факт, который не наносит ему лично ущерба, но противоречит его представлениям о справедливости или этическим нормам].

2) Хотя оба синонима предполагают, что субъект поступает вопреки своим интересам и приносит те или иные жертвы, в случае *благородный* они обычно более значительны, чем в случае *великодушный*.

Во-первых, *благородный* часто указывает на то, что субъект без всякой личной заинтересованности совершает действия (или осуществляет деятельность), требующие больши’х усилий и сопряженные с риском – борьба с кем-л., спасение кого-л. и пр.; ср. *Благородный герой спасает девушку <борется за справедливость, вступается за слабых>*. Между тем, в случае *великодушный* субъект чаще всего просто воздерживается от каких-л. действий или совершает действия, не требующие больши’х затрат сил – дает что-л., предоставляет возможность сделать что-л. и пр., причем в большинстве случаев эти действия носят речевую форму[4]. Отчасти это связано с тем, что *великодушный* предполагает власть субъекта над партнером по общению и над ситуацией в целом, поэтому для достижения цели субъекту не приходится прикладывать много усилий – ему обычно достаточно совершить речевое действие (*предложить, разрешить, пригласить* и пр.).

Во-вторых, *благородный* допускает большой масштаб ущерба, который понес субъект, что для *великодушный* нетипично. Ср. *Благородный – пожертвовал всем ради нее; Разве это не благородно - защитить нас своими телами?* (Д. Емец, Таня Гроттер и колодец Посейдона); *Бы́ло в той де́тской стра́стности и гото́вно-*

[4] Ср. более естественное *великодушно предложил <пообещал> помочь* при менее естественном великодушно помог.

040-Krylova_T:_Layout 1 18.05.2008 15:30 Page 265

сти отда́ть жизнь за далёкого чужо́го что́-то от ева́нгельского «за други своя» неподде́льное и благоро́дное (А.Варламов, Купавна). Слово *великодушный* в этих контекстах выглядело бы менее естественно.

3) Как уже говорилось, оба синонима могут описывать ситуацию, когда субъект, вопреки своей выгоде, не делает чего-л. плохого. Однако *благородный,* в отличие от *великодушный,* применим не только в ситуации, когда негативные действия носят закономерный характер и являются логическим следствием отношений между субъектом и его партнером (или поступков последнего). Он может использоваться и тогда, когда субъект воздерживается от негативных действий, которых от него не ожидали, поскольку они противоречат характеру отношений между участниками общения. Ср. нормальное *Он не предал нас – поступил благородно*; *Благородный – никогда не совершит подлость;* слова *великодушно, великодушный* здесь были бы неуместны.

4) *Благородный* не указывает на преимущество субъекта перед партнером по общению.

5) *Великодушный* обозначает свойство, которое может приобретаться (*Успех сделал его великодушным; С возрастом он стал великодушнее*) и контролироваться человеком (*Будь великодушным; Нужно быть великодушным; В этой ситуации он мог позволить себе быть великодушным*). Между тем, *благородный* обозначает свойство, обычно присущее человеку от рождения (ср. *врожденное благородство*) и не подвластное человеческой воле – на это указывает неподставимость этого слова и его дериватов в вышеприведенные контексты.

Таким образом, на наш взгляд, *благородный* обозначает более редкое (и ценное) свойство, чем *великодушный. Благородный* человек способен на бо'льшие жертвы, его доброта и милосердие не обусловлены его преимуществом; кроме того, это свойство, в отличие от свойства быть *великодушным*, носит врожденный характер и более универсально – в частности, допускает в качестве варианта бескорыстное служение идее.

Внутренняя форма слова *благородный* ('хорошего рода') отсылает нас к уходящему значению этого прилагательного, которое реализуется в сочетаниях типа *благородные господа.* Именно от него было произведено интересующее нас 'этическое' значение прилагательного *благородный*, которое реализуется в современной лексеме *благородный 1.1.* Изначально *благородный* в его этическом значении должно было включать в свое толкование элемент как 'такой, который ведет себя так, как это свойственно человеку, имеющему высокое происхождение'.

Такой механизм переноса наводит на мысль, что на более раннем этапе развития языка, в социальную эпоху, когда фактор происхождения был более актуален для русской культуры и общества, в наивно-языковом сознании существовало представление о связи между *благородством* как этическим качеством и социальным происхождением человека[5].

Следы сходных представлений о связи между этическими свойствами и социальным происхождением мы находим в структуре многозначности существительного *хам*, которое, наряду с основным значением ('человек, который сознательно нарушает нормы уважительности') имеет уходящее значение 'человек низкого происхождения'. Ср. *Это безразлично. Мастеровой, рабочий - одно и то же. По-ихнему - рабочие, а по-нашему – хамы* (В. Катаев, Миллион терзаний); *Да куда вы? Помилуйте, ведь опасно. Теперь за каждым углом караулят. Как из нашей зоны выйдете, сейчас вас схватят хамы*, *а то и подстрелят без разговору* (А. Серафимович, Две смерти); *Понимаете? Никаких суперфосфатов, азотистых солей, селитры! Удобрить поля миллионами! Мужичье, хамы*, *взбунтовавшаяся сволочь. Все в машину! Большую кофейную мельницу. В кашу!* (Б. Лавренев, Рассказ о простой вещи).

При этом следует отметить, что в случае *благородный* первичным было значение, характеризующее происхождение, тогда как в случае *хам* исходным следует признать значение, характеризующее этические свойства.

В случае *благородный,* в связи с изменениями в обществе, вызвавшими устаревание соответствующей лексемы в значении 'имеющий высокое происхождение', основной стала лексема с этическим значением.

Интересно, что некоторое представление о связи между социальным происхождением и внутренними свойствами человека сохраняется в наивно-языковом сознании до сих пор, и об этом свидетельствует не только тот факт, что *благородный 1.1* и *2.1* (*благородные господа*) мы осмысляем не как омонимы, а как разные лексемы в рамках единого многозначного слова.

Во-первых, на это указывает представление о наследственном, врожденном характере этического свойства в случае *благородный 1.1*, о котором шла речь выше.

Во-вторых, об этом свидетельствует значение ряда лексем вокабулы *благородный.*

Чтобы проиллюстрировать это, приведем вкратце схему многозначности этого прилагательного.

Благородный 1.1 'такой, который руководствуется в своих поступках не своими личными интересами, а интересами других людей или отвлеченными представлениями о том, что хорошо, а что плохо'; *благородные борцы за справедливость*; *Благоро́дный, му́жественный и немногосло́вный геро́й спаса́ет обречённого пса* (А.

[5] Можно попытаться восстановить логику таких представлений: коль скоро способность к альтруизму отличает человека от животного, человеческая «селекция» должна приводить к удалению от «животного» начала с его эгоизмом и заботой о собственных интересах и к развитию альтруизма.

040-Krylova_T:_Layout 1 18.05.2008 15:30 Page 266

Геласимов, Фокс Малдер похож на свинью).

Благородный 1.2 ‘такой, которого естественно ожидать от *благородного 1.1*’; *благородная задача <цель, миссия, попытка>; благородный жест <порыв>; благородные идеалы; благородная тяга к справедливости.*

Благородный 2.1 уходящ. ‘такой, который имеет высокое происхождение’; *благородные господа, благородная дама, благородный рыцарь; благородное происхождение, благородных кровей; Взял ба́бу-то из благоро́дных, тепе́рь отдува́ться прихо́дится...*(Л. Улицкая, Пиковая дама); *У меня́ роди́тели вполне́ благоро́дные лю́ди из го́рода Профцгейма, а я их зако́нный сыно́к!* (Ю.Домбровский, Обезьяна приходит за своим черепом); ср. также номинации *Благородное собрание, Институт благородных девиц.*

Благородный 2.2 [о видах объектов] ‘такой, который встречается редко и высоко ценится’; *благородный сорт, благородная порода; Севрюга – благородная рыба*; ср. также название растения *лавр благородный.*

Благородный 3.1 [о внешности человека] ‘такой, который бывает у людей, имеющих высокое происхождение, или по которому видно, что человек обладает высокими духовными или нравственными качествами’; *благородные руки, благородный лоб <профиль>, благородное лицо, благородная внешность, благородный облик.*

Благородный 3.2 [обычно в сочетании со словом *животное*] ‘такой, во внешности или поведении которого есть что-то величественное, и который отличается гордостью, смелостью и независимостью’; *Кошки <львы, лошади, олени> – благородные животные; Она обняла за шею благородное животное; А так как среди́ звере́й счита́лся царём лев, то есть са́мый си́льный, са́мый хра́брый и са́мый благоро́дный зверь, то я, есте́ственно, счита́л, что лю́ди в вы́боре своего́ царя́ по́льзуются не менее разу́мными при́знаками* (Ф. Искандер, Дедушка); *Благородное животное поднялось на дыбы, сделав эффектную свечку, и с места взяло в галоп* (А. Белянин, Свирепый ландграф).

Благородный 4 [о внешних параметрах или материальных объектах] ‘такой, в котором нет ничего лишнего, чрезмерно привлекающего внимание или нарушающего гармонию’; *благородный цвет, благородная форма <красота>, благородная ткань, благородные линии, благородный тембр <запах>; благородный дом <памятник>; Там стои́т о́чень благоро́дный, краси́вый па́мятник - бе́лая мра́морная плита́ и на ней про́филь Владислава Игнатьевича* (С. Спивакова, Не всё).

Рассмотрим лексему *благородный 3.1.* Она имеет два режима употребления.

1) В сочетании с названиями таких деталей внешности, как *профиль, овал, руки, осанка, посадка головы,* лексема *благородный 3.1* значит ‘такой, который бывает у людей, имеющих высокое происхождение’. В сочетаниях *благородный профиль <овал>, благородные руки* лексема *благородный* обычно характеризует форму частей тела или лица, указывая на их удлиненность, тонкость и изящество; ср. *благородные руки* [с тонкими, длинными пальцами]*; благородный профиль* [удлиненный, с тонким носом], *благородный овал лица* [вытянутый]. Ср. *Алые народовольческие пятна залили все бледное гневное лицо и благородные руки, испарина выступила на аттическом лбу* (Т. Устинова, Подруга особого назначения). В сочетаниях *благородная осанка <посадка головы>* рассматриваемая лексема указывает на величественность, горделивость.

2) В сочетании со словом *лоб* и книжной лексемой *чело* прилагательное *благородный* обычно значит ‘такой, по которому видно, что человек обладает высокими духовными или нравственными качествами’; ср. *благородное чело; Если прибавить, что лоб Льва Николаевича с достаточно развитыми лобными буграми и достаточной высоты, слегка превышающей длину носа, то можно его лоб характеризовать так: красивый, благородный лоб мыслителя* (В. Люстрицкий, Лев Николаевич Толстой в Московской окружной лечебнице для душевнобольных).

В сочетании со словом *лицо* у лексемы *благородный* могут реализоваться оба рассмотренных подзначения: при наличии прилагательных, характеризующих внешний вид (*удлиненное <узкое, утонченное> благородное лицо, сухое благородное лицо*) реализуется первое из них, между тем, в контексте прилагательных, характеризующих внутренние свойства субъекта, реализуется второе; ср. *открытое благородное лицо; честное, благородное лицо; Какое благородное лицо у старика: смесь незлобия и проницательности* (Н. Г. Чернышевский, Что делать?); *Лица под неоновым светом казались благороднее и чище* (С. Довлатов, Иная жизнь).

В некоторых случаях границу между двумя рассматриваемыми режимами употребления провести достаточно трудно; ср. *гордое и благородное лицо Дон Кихота; Лысина не портила его благородного облика; Она стала на ощупь считать зеленые пятерки, вызывая в памяти благородное лицо Линкольна* (В. Пелевин, Миттельшпиль); *Лицо на зависть благородное. Такое ощущение, что Гарвард закончил* (С. Довлатов, Переводные картинки).

Совмещение этих двух подзначений в рамках одной лексемы позволяет говорить о том, что в современном языковом сознании до сих пор сохраняется представление о связи между происхождением и внутренними качествами человека.

Похожим образом устроено значение *благородный 3.2,* которое реализуется главным образом в сочетании со словом *животное.* Соответствующая лексема одновременно характеризует внешние и внутренние качества.

040-Krylova_T:_Layout 1 18.05.2008 15:30 Page 267

С одной стороны, она указывает как величественную, царственную внешность (которая сближает *благородное животное* с человеком *благородного* происхождения). Не случайно *благородными* обычно называют только животных достаточно большого размера, «вертикальной ориентации», с длинной шеей, которая создает впечатление взгляда «сверху вниз» (бегемота, несмотря на большой размер, не назовешь *благородным* животным), и медлительными движениями. С другой стороны, когда мы называем животное *благородным,* мы имеем в виду еще и то, что оно наделено определенными внутренними качествами[6] - прежде всего теми, которые в мире людей ассоциируются с высоким происхождением – гордостью, независимостью, смелостью, отчасти – благородством (в данном случае речь идет об отсутствии бессмысленной жестокости и коварства). При этом в разных ситуациях на первый план могут выходить разные внутренние качества: ср. немного разное значение прилагательного *благородный* в следующих контекстах: *Олень – благородное животное* и *Лев – благородное животное.*

Заслуживает отдельного внимания также последняя лексема *благородный 4,* которая указывает на соответствие материального объекта или его параметров требованиям хорошего вкуса, на отсутствие в нем вульгарности и пошлости. Учитывая, что все значения прилагательного *благородный* развились из значения ‘такой, который имеет высокое происхождение’, наличие у этого слова лексемы, подобной *благородный 3.2,* свидетельствует о том, что когда-то в наивно-языковом сознании существовали также представления о связи между эстетическими предпочтениями человека и его происхождением. Если толпе нравится кричащее, броское, то «элита» предпочитает нечто сдержанное и неброское (ср. также аналогичное «вкусовое» значение слова *вульгарный,* развившееся на основании исходного значения ‘народный’). Впрочем, в современном языке смысловая связь между рассматриваемыми значениями уже практически не ощущается.

[6] В отличие от *благородного 3.1*, в случае *благородный 3.2* оба этих компонента присутствуют одновременно.

041-Kruchkova_OJu:_Layout 1 18.05.2008 15:35 Page 268

ТЕКСТОВЫЙ ДИАЛЕКТОЛОГИЧЕСКИЙ КОРПУС КАК МОДЕЛЬ ТРАДИЦИОННОЙ СЕЛЬСКОЙ КОММУНИКАЦИИ[1]

TEXTUAL DIALECT CORPUS AS A MODEL OF TRADITIONAL RURAL COMMUNICATION

Крючкова О.Ю. *(vpks@rambler.ru),* ***Гольдин В.Е.****(goldinve@yandex.ru)*
Саратовский государственный университет им. Н.Г. Чернышевского

В докладе обсуждаются принципы организации и методика построения мультимедийного диалектологического текстового корпуса, представляющего диалект как целостное культурно-коммуникативное образование, моделирующего коммуникацию конкретных речевых коллективов в конкретных социокультурных условиях.

1. Введение

В лингвистике уже разработаны принципы построения текстовых корпусов как коммуникативных моделей (Британский национальный корпус, Национальный корпус русского языка и др.). Наряду с такими корпусами, представляющими функционирование национальных языков в различных сферах общения, необходимы также и корпуса, моделирующие коммуникацию в отдельных языковых сообществах, выделяемых в рамках национального языка. Важнейшими языковыми образованиями такого типа являются диалекты.

Создание диалектных текстовых корпусов имеет особую значимость для диалектологии, ограниченной до последнего времени кругом источников. Основным источником диалектологии до недавнего времени оставался материал, собранный по специальным вопросникам и представленный в картотеках, словарях, атласах. Специфика материала определяла и особый, периферийный, статус диалектологии в составе русистики, значительно сужала круг решаемых диалектологией задач, ограничивая диалектологическую проблематику в основном сферой системно-структурного своеобразия русских народных говоров. Первичный же диалектный материал, имеющийся в различных диалектологических центрах, во-первых, все еще мало доступен широкому кругу исследователей, а во-вторых, в том виде, в котором он обычно существует (без специального аннотирования) не может быть использован с максимальной пользой.

Корпуса диалектной речи, отражающие коммуникацию на диалекте в том или ином конкретном населенном пункте и сохраняющие в машинообрабатываемой форме значительные массивы связной речи, являются основным источником изучения коммуникативной специфики диалектов. Использование материалов таких корпусов дает возможность не ограничиваться отдельными примерами, а переходить к выявлению общих принципов, тенденций, действующих в диалектной коммуникации (см., напр. работы, выполненные на материале СДК [Гольдин, Крючкова 2006; Балаян 2006; Крючкова, Гольдин, Сдобнова 2007; Крючкова 2007а; Крючкова, Гольдин 2007; Крючкова, Сдобнова, Гольдин 2007; Свешникова 2007а; Свешникова 2007б]), позволяет составить целостное представление о специфике традиционного сельского общения, построить симметричное литературному языку описание диалектной речи.

2. Характеристика имеющихся диалектных корпусов

Мысль о необходимости создания машинного фонда диалектных текстов была высказана еще в 1980-х г. А.С. Гердом и аргументирована В.Е. Гольдиным [Машинный фонд, 1986: 72; Гольдин, 1990]. Однако до сих пор в отечественной и зарубежной лингвистике отсутствуют крупные корпуса диалектной речи, моделирующие коммуникацию на диалекте.

В настоящее время существует ряд корпусов, репрезентирующих отдельные элементы диалектной речи: зарубежные корпуса диалектных текстов (например, Helsinki corpus of English dialects, Kirk's Northern Ireland Transcribed Corpus of Speech (NITCS), IViE (Intonational Variation in English) corpus, BBC Voices); диалектный подкорпус в составе Национального корпуса русского языка (НКРЯ); лексико-грамматическая база данных (ЛГБД) по говору с. Пустоша Шатурского р-на Московской обл., включающая тексты – образцы речи носителей говора [Тер-Аванесова, Крылов, 2006].

[1] Работа выполнена при поддержке Российского фонда фундаментальных исследований (РФФИ), проект № 06-06-80428-а

041-Kruchkova_OJu:_Layout 1 18.05.2008 15:35 Page 269

Основная задача диалектного подкорпуса НКРЯ заключается в представлении диалекта как специфической территориальной разновидности общенародного языка. Диалектный подкорпус НКРЯ включает текстовые фрагменты различных говоров (напр., говоры Архангельской области, Волгоградской, Рязанской, Вологодской областей), выделенные на тематической основе (быт, гадание, обычаи, жизнь и др.). Методической основой корпуса является последовательное сравнение диалекта с литературным языком (прежде всего в области морфологии и лексики): «в диалектном подкорпусе специально отмечаются отличия от литературного языка» [Летучий, 2005: 215]. Для этой цели при разметке корпуса используется ряд дифференциальных помет, фиксирующих параметры, по которым диалектная текстоформа отличается от соответствующей ей литературной формы (пометы dialmorph, diallex с их последующей конкретизацией); подробнее о диалектном подкорпусе НКРЯ в сопоставлении с Саратовским диалектологическим корпусом см. [Крючкова 2007б].

ЛГБД по говору с. Пустоша отличается ориентацией на один конкретный говор и имеет целью «исчерпывающее описание говора в рамках определенного корпуса текстов» [Тер-Аванесова, Крылов, 2006]. Аннотированный диалектный корпус говора с. Пустоша дает сведения о фонетической, морфологической и лексической специфике текстовых словоформ говора. В корпусе, как и в НКРЯ, использована сравнительная методика описания диалекта, при которой диалектная специфика рассматривается в качестве «отклонений» от литературного аналога.

Таким образом, имеющиеся диалектологические корпуса объединяет дифференциальный подход к диалекту, направленный на выделение элементов, отличающих диалект от литературного языка. Диалектологические корпуса, построенные на дифференциальной основе, демонстрируют территориальное варьирование национального языка, но дают ограниченное представление о традиционном сельском (диалектном) общении как о целостном культурно-коммуникативном феномене.

3. Саратовский диалектологический корпус (СДК)

3.1. Принципы организации корпуса

В Саратовском государственном университете им. Н.Г. Чернышевского создается мультимедийный диалектологический текстовый корпус, целью которого является презентация диалекта как целостного культурно-коммуникативного образования, построение модели традиционного сельского общения на диалекте. СДК базируется на представлении о диалекте как о самодостаточной коммуникативной системе, полно обеспечивающей коммуникативные потребности в условиях традиционного сельского общения. При создании корпуса реализуется недифференциальный подход к диалекту, при котором рассмотрение диалектной речи в ее отношении к литературному языку не является основной задачей.

Каждый отдельный говор в СДК образует самостоятельный подкорпус. В настоящее время корпус включает 3 самостоятельных подкорпуса: подкорпус говора с. Белогорное Вольского района Саратовской области, подкорпус говора с. Земляные Хутора Аткарского района Саратовской области и подкорпус куста сел Мегра Вытегорского района Вологодской области (состав подкорпусов может увеличиваться).

Задача создания модели традиционной сельской коммуникации требует разработки специальной программы, обеспечивающей репрезентативность включаемого в корпус материала. Общим принципом формирования текстовой базы корпуса является принцип полного и адекватного отражения в корпусе специфики диалектного общения. Реализация данного принципа предполагает наполнение каждого подкорпуса разнообразным значительным по объему текстовым материалом, репрезентирующим:

– важнейшие типы диалектной речи (речь бытовую, фольклорную, речь в условиях официального, обрядового общения);

– различные формы речи (диалог, полилог, монолог);

– разнообразную тематику сельского общения;

– социальную дифференциацию носителей говора (по полу, возрасту, профессии, уровню образования).

Для построения коммуникативной модели речевого общения в конкретных условиях жизни конкретного речевого коллектива необходим учет социокультурных условий бытования говора. Решение этой задачи осуществляется путем включения в состав каждого подкорпуса многообразной нелингвистической информации: фотографий, видеоиллюстраций, схем, карт, сведений исторического, социокультурного характера, демографических, этнографических, географических данных. Часть данной информации соотнесена в корпусе с текстовыми модулями, другая часть образует отдельный информационный блок (см. Рис. 1).

041-Kruchkova_OJu:_Layout 1 18.05.2008 15:35 Page 270

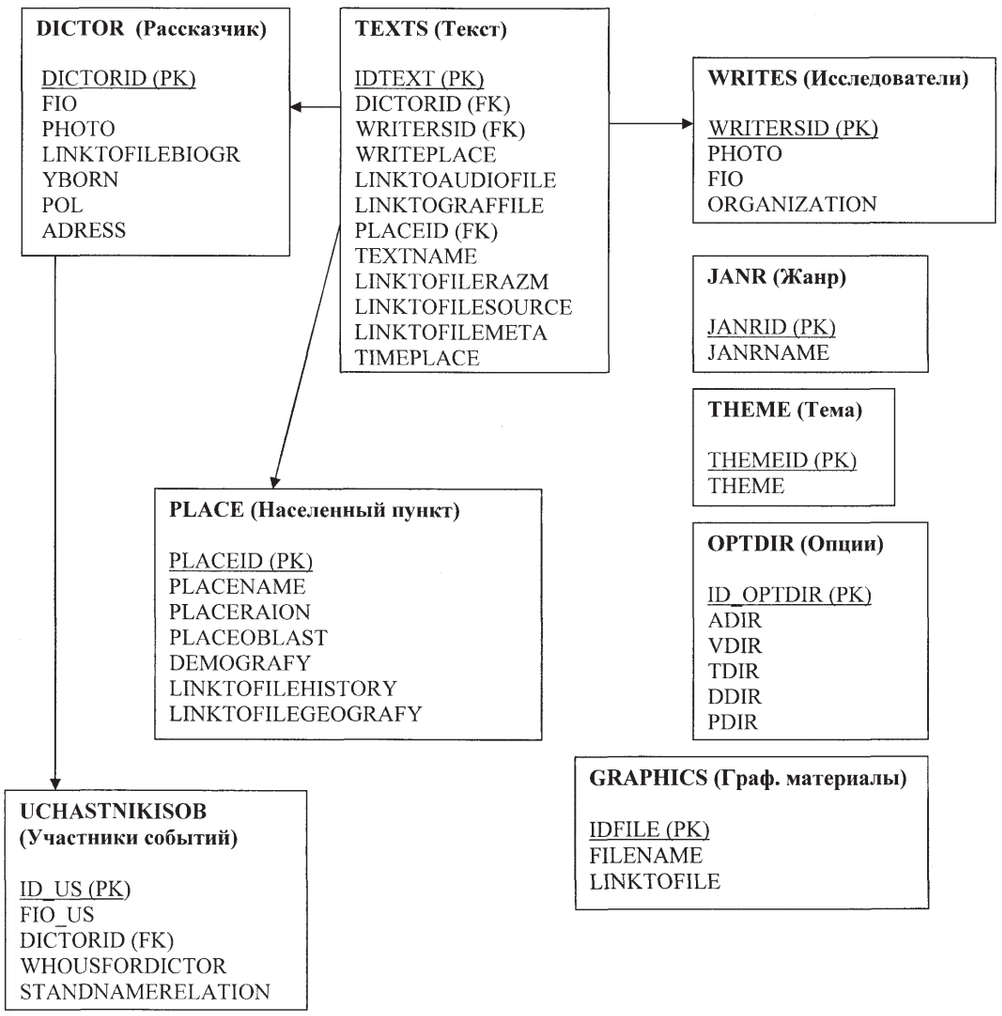


Рис. 1. Информационные блоки СДК

Мультимедийная организация электронной базы корпуса позволяет получать на выходе как отдельные кванты содержащейся в корпусе информации, так и информацию комплексного характера, переходить от одного информационного блока к другому.

3.2. Представление в корпусе диалектной речи

В отношении текстовой (основной) части диалектологического корпуса необходимо решение следующих принципиальных вопросов:

1) способ хранения диалектной речи в базе корпуса;

2) способ членения потока речи;

3) способ символьной записи (расшифровки) устной диалектной речи.

Решения, принятые в СДК:

1) Основная единица базы СДК – «запись»/ «текст» – хранится в трех видах: в виде звукового модуля, в виде текстового модуля с символьной расшифровкой аудиозаписи и в виде текстового модуля с размеченным текстом. От текстовых модулей возможен переход к звуковым модулям и наоборот; программное обеспечение корпуса дает также возможность одновременного воспроизведения аудиозаписи, символьной расшифровки диалектного текста и размеченного текста. В последнее время архив СДК пополняется видеозаписями диалектной коммуникации, значимость которых велика ввиду наиболее полной фиксации коммуникативной ситуации и соотношения между вербальным и невербальным компонентами общения. При наличии видеозаписи звуковой модуль может быть заменен видеомодулем, также программно связанным с текстовыми модулями.

041-Kruchkova_OJu:_Layout 1 18.05.2008 15:35 Page 271

2) Проблема членения речевого потока решена в корпусе в соответствии с принципом максимального приближения модели прототипическому объекту – естественной коммуникации на диалекте. Границы «записи»/«текста» определяются с помощью формального критерия непрерывности общения, так что речевые фрагменты аудио-/видеофайла и текстовых файлов полностью совпадают и соответствуют зафиксированному звукозаписывающей и видеоаппаратурой непрерывному фрагменту общения. В результате применения данного критерия границы звуковых и текстовых модулей не зависят от таких параметров, как смена темы, жанра, формы речи, изменение коммуникативной ситуации и числа участников коммуникации.

3) Расшифровки звучащей диалектной речи приводятся в близкой к орфографической символьной записи, с отражением лексических и грамматических особенностей диалектной речи. Регулярные фонетические явления (например, характер безударного вокализма, диалектные различия в области консонантизма, такие, как диалектное произношение /г/ или /л/) в символьных расшифровках не отображаются. Отражение получают лишь лексикализованные фонетические особенности (типа *теперь, кстить, Рожество, Паска*).

Пример символьной расшифровки диалектного текста в СДК:

вот это у нас образа/ оне называются старообрядческие/ поморские/ вот наша вера какая/ мы не монашки/ мы не какие поповцы/ и не это… вот оне эти вот/ монашки-ти/ оне… как сказать… оне… тоже беспоповцы/ но… у них брак/ и считается за грех/ оне называются безбрачные// оне вот доживают/ до шестьдесят лет/ женщина/ если с мужчиной живёт/ это было раньше/ и у них закон такой/ доживают до шестьдесят лет/ она сносит кануны/ и говорит/ вот/ Ваня там или Вася/ всё/ я больше прекращаю с тобой жизню жить/

Отсутствие фонетической транскрипции в символьных записях восполняется включением в СДК звуковых модулей. Параллелизм текстовых модулей и аудио-/видеомодулей обеспечивает максимальную достоверность фонетической информации, возможность ее наиболее объективного использования в диалектологических исследованиях.

Символьная расшифровка звучащей речи требует также решения целого ряда частных проблем, таких, как возможность и характер использования знаков препинания, обозначение неразобранных фрагментов речи и недоговоренных слов, дифференциация речевых отрезков, принадлежащих диалектологу и диалектоносителю, а также разным диалектоносителям, способ подачи необходимых для понимания текста комментариев. Принятые в связи с данными вопросами решения зафиксированы в специальной инструкции по расшифровке и разметке диалектных текстов СДК (инструкция размещена на сайте www.sarteorlingv.narod.ru).

3.3. Многоуровневая разметка текстов и типы программно обрабатываемой текстовой и метатекстовой информации

Аннотирование текстовой базы корпуса проводится путем многоуровневой лингвистической параметризации диалектных текстов и их метаописания, представляемого в отдельных файлах.

Виды осуществляемой в СДК разметки символьных расшифровок диалектных текстов:

1) пословная лексико-морфологическая разметка;

2) жанровая разметка;

3) тематическая разметка.

Характер разметки определяется реализованным в СДК недифференциальным подходом к диалекту и регулируется следующими положениями:

– все бытующие в говоре языковые формы (совпадающие и не совпадающие с литературными) являются элементами диалектной языковой системы данного говора;

– устное диалектное повествование носит принципиально политематический характер и отличается нечеткими тематико-жанровыми границами.

Названные положения обусловливают особенности проводимой в СДК лексико-морфологической, жанровой и тематической разметки.

Пословная лексико-морфологическая разметка в СДК, полно описывающая все морфологические (классификационные и словоизменительные) признаки текстовой формы и семантику лексических диалектизмов, во многом опирается на принципы, выработанные при разметке текстов в НКРЯ. Разметка проводится с помощью автоматического анализатора с последующим ручным редактированием. Вместе с тем недифференциальный характер СДК обусловил ряд отличий при лексико-морфологической разметке текстов: отказ от дифференциальных помет, применяемых в НКРЯ; характер подачи начальной формы; введение зоны литературных соответствий.

Отказ от дифференциальных помет при пословной лексико-морфологической разметке в СДК не исключает, однако, возможности поиска текстоформ, отличающихся по какому-либо признаку от литературных

041-Kruchkova_OJu:_Layout 1 18.05.2008 15:35 Page 272

словоформ. Возможность такого поиска обеспечивается специальной маркировкой не соответствующих литературной норме единиц знаком «*», помещаемым в дополнительной зоне разметки.

Начальная форма восстанавливается в СДК на основе конкретной текстоформы, например, для текстоформы *ходилась* приводится начальная форма *ходиться*, для *посклизнулся – посклизнуться*, для *куды – куды* и т.д.

Отказ от дифференциальных помет и текстоориентированная лемматизация словоформ, безусловно, затруднили бы поиск в текстовой базе СДК. Для облегчения поисковых запросов в лексико-морфологическую разметку вводится зона литературных соответствий, идущая в разметке вслед за начальной формой. Литературное соответствие при грамматических, лексических и словообразовательных диалектизмах выполняет также функцию семантической перекодировки. Информативным является и отсутствие литературного соответствия, причиной которого могут быть неясность значения нелитературного слова либо языковая лакуна. В первом случае в зоне литературного соответствия ставится знак вопроса, во втором – тире (прочерк). Ср.: *бедранку* {бедранка(?)=S,жен,неод=ед,вин} (*и он вот под бедранку-то/ залез*); *дак* {дак(–)=PART}, и {и(–)=PART} в функции маркеров конца высказывания.

Примеры пословной лексико-морфологической разметки в СДК:

оне {оне(*они*)=S,мн,од=им=*};

сварывать {сварывать(*сворачивать*)=V,несов=инф=*};

цементовый {цементовый(*цементный*)=A=ед,муж,им=*}.

Ввиду нечеткой выделимости в диалектной коммуникации целостных в жанровом и тематическом отношениях отрезков речи и многочисленных жанрово-тематических переходов и наложений жанровая и тематическая кодировки диалектного текста проводятся на основе предельно обобщенной (не конкретизированной) рубрикации. Элементами жанровой параметризации являются, например, «рассказ-повествование», «рассуждение», «описание», «сказка», «песня», «пословицы, поговорки», а элементами тематической рубрикации – такие темы, как «семья», «обряды. обычаи, приметы», «здоровье и лечение», «религия», «природа», «происшествия».

С каждым текстом связаны 3 модуля метатекстового характера:

1) модуль с метаразметкой текста, элементами которой являются сведения об информантах, о времени и месте записи текста, о конкретной ситуации общения, об адресатах речи, об упоминаемых в тексте лицах, о времени описываемых в тексте событий (до революции; революция и гражданская война; коллективизация; Великая отечественная война; послевоенный советский период; постсоветский период), перечень представленных в тексте тем и перечень жанров;

2) модуль, содержащий биографию информанта, восстанавливаемую по текстовым данным;

3) иллюстративный модуль (фотографии информанта, фотоиллюстрации к данному тексту).

Таким образом, обработка каждого текстового модуля для включения его в СДК завершается формированием папки со следующим набором разноформатных модулей: аудио- / видеомодуль, 4 текстовых модуля (символьная расшифровка аудиозаписи, размеченный диалектный текст, метаописание текста, биография информанта), 1 иллюстративный модуль.

Все элементы пословной лексико-морфологической разметки, жанровой и тематической кодировки диалектных текстов, а также метатекстовая информация являются параметрами поисковых запросов и образуют программно связанное целое.

4. Заключение

Образуя содержательно и программно связанное мультимедийное средство, СДК дает возможность получать комплексную информацию о говоре и условиях его бытования. Электронный диалектологический корпус, моделирующий традиционную сельскую коммуникацию на диалекте, является принципиально новым источником изучения диалектной речи, соответствующим современным требованиям науки о русских народных говорах. Создание СДК и других текстовых диалектологических корпусов будет способствовать приобщению диалектологии к современной научной лингвистической парадигме и построению такой русистики, в которой изучение основных языковых страт (литературной и диалектной речи) находилось бы в необходимой и корректной корреляции.

Список литературы

1. Балаян Э.В. Межконфессиональные отношения в с. Белогорном Вольского района Саратовской области по рассказам местных жителей // Народы Саратовского Поволжья: этнология, этнография, духовная и материальная

041-Kruchkova_OJu:_Layout 1 18.05.2008 15:35 Page 273

культура: Материалы межрегиональной научно-практической конференции. Труды Саратовского областного музея краеведения. Саратов: СОМК. 2006. № 10.

2. Гольдин В.Е., Крючкова О.Ю. Тематическая разметка и тематический анализ диалектного текстового корпуса // Языковая личность – текст – дискурс: теоретические и прикладные аспекты исследования: материалы междунар. научн. конф.: в 2-х ч. Ч.1. Самара: Изд-во «Самарский университет», 2006.

3. Гольдин В.Е. К проекту текстового диалектологического подфонда Машинного фонда русского языка // Доклады Третьей Всероссийской конференции по созданию Машинного фонда русского языка. М., 1990.

4. Крючкова О.Ю. Оценки речи как проявление культурно-языковой идентичности носителей диалекта // Проблема идентичности в современном мире. Саратов: Изд-во Саратовского университета, 2007а.

5. Крючкова О.Ю. Электронный корпус русской диалектной речи и принципы его разметки // Известия Саратовского университета. Новая серия. Филология. Журналистика. Саратов: Изд-во Саратовского университета, 2007б. Т. 7. Вып. 1.

6. Крючкова О.Ю., Гольдин В.Е. Морфологическое своеобразие среднерусской диалектной речи по данным диалектных текстов // Текст и языковая личность: Материалы V Всероссийской научной конференции с международным участием. Томск: Изд-во ЦНТИ, 2007.

7. Крючкова О.Ю., Гольдин В.Е., Сдобнова А.П. Электронный диалектный корпус как новый источник изучения русских народных говоров // Язык и культура в России: состояние и эволюционные процессы: материалы международной научной конференции. Самара: Изд-во «Самарский университет», 2007.

8. Крючкова О.Ю., Сдобнова А.П., Гольдин В.Е. Лексическое своеобразие среднерусской диалектной речи по данным диалектного текстового корпуса // Античный мир и мы: Межвузовский сборник научных трудов. Саратов: Изд-во Саратовского мед. ун-та. 2007. Т.2. № 11.

9. Летучий А.Б. Корпус диалектных текстов: задачи и проблемы // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. М., 2005.

10. Машинный фонд русского языка: идеи и суждения. М., 1986.

11. Свешникова Н.В. Модели диминутивного словообразования в русских говорах (функциональный аспект) // Диалектное словообразование, морфемика и морфонология. СПб.: Наука; Вологда: ВГПУ, 2007а.

12. Свешникова Н.В. Явление лексикализации фонетических особенностей в одном из среднерусских говоров // Язык и культура: Материалы Международной научной конференции, посвященной 70-летию профессора Л.В. Савельевой. Петрозаводск: Изд-во КГПУ, 2007б.

13. Тер-Аванесова А.В., Крылов С.А. Лексико-грамматические базы данных как инструмент диалектологического описания // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции «Диалог 2006». М.: Изд-во РГГУ, 2006.

042-Kudashev_I:_Layout 1 18.05.2008 15:35 Page 274

ПОЛЕЗНЫЕ ДОПОЛНЕНИЯ К ТРАДИЦИОННОЙ ПРАКТИКЕ СОСТАВЛЕНИЯ ПЕРЕВОДНЫХ ТЕРМИНОЛОГИЧЕСКИХ СЛОВАРЕЙ (НА ПРИМЕРЕ ДВУХ ФИНСКО-РУССКИХ СЛОВАРЕЙ)

USEFUL EXTENSIONS TO TRANSLATION-ORIENTED TERMINOLOGICAL DICTIONARIES (CASE: TWO FINNISH-RUSSIAN DICTIONARIES)

Кудашев И.С. *(igor.kudashev@helsinki.fi),* ***Кудашева И.О.*** *(irina.kudasheva@helsinki.fi)*
Хельсинкский университет, Финляндия

В статье описывается ряд полезных дополнений к традиционной практике составления переводных терминологических словарей. Указанные приемы использовались при разработке двух переводных терминологических словарей, составленных в 2003–2007 гг. в Хельсинкском университете.

В отличие от многих других направлений лексикографии практика составления терминологических словарей за последние 20–30 лет не претерпела существенных изменений. Во многих европейских странах за образец приняты стандарты ИСО, такие как ISO 704:2000 (Terminology Work – Principles and Methods) и ISO 12620:1999 (Computer Applications in Terminology – Data Categories), а также основанные на них руководства по терминологической работе. В Советском Союзе, а затем в России, по наблюдениям В.А. Татаринова (1996: 141), оформление терминологических словарей также приобрело со временем такой уровень унификации, что исследования в этой области стали как бы излишними.

В данной статье описываются некоторые дополнения к традиционной практике составления переводных терминологических словарей, которые мы нашли полезными при работе над двумя терминологическими словарями, составленными в 2003–2007 гг. в Центре обучения и развития «Палмения» Хельсинкского университета. Речь идет о «Финско-русском лесном словаре» (2008) и англо-финско-русском «Словаре лексики проектного сотрудничества ЕС – Россия» (2008). Указанные словари были составлены авторами данного доклада при финансовой поддержке программы соседства «Юго-Восточная Финляндия – Россия», Интеррег III А и ряда финских государственных, муниципальных и общественных организаций. Научными консультантами и редакторами обоих словарей выступали профессор Хельсинкского университета Инкери Вехмас-Лехто и профессор Санкт-Петербургского университета А.С. Герд. Словари создавались в сотрудничестве с десятками российских и финских специалистов-предметников. Мы надеемся, что некоторые приемы, использованные в этих словарях, заинтересуют других составителей и со временем получат более широкое распространение в терминографической практике.

Сочетание общей дескриптивной направленности словаря с элементами нормативности

В терминографии четче, чем в других направлениях лексикографии, различаются языковая и организационная норма. Во многих странах существуют терминологические центры, имеющие полномочия регламентировать терминоупотребление в специальных сферах, главным образом путем издания государственных и межгосударственных стандартов на термины и определения. Эти издания носят нормативный характер, и в них используется развитая система предписывающих помет, таких как исп. вместо (используй вместо), нерек., нрк. (нерекомендуемый термин), ндп. (недопустимый термин). Нормативные словари обычно отличаются небольшим объемом, и значительная часть нерекомендуемых терминов в них просто не попадает.

Словари, составляемые вне терминологических центров, как правило, носят дескриптивный характер и стремятся к отражению всех лексических пластов языков для специальных целей, однако в них редко встречаются оценки лексических единиц с точки зрения языковой правильности и языковых конвенций, принятых в той или иной предметной сфере.

При составлении наших словарей мы стремились совместить преимущества дескриптивного и нормативного подхода. Лексические единицы включались в словари без оглядки на степень их нормативности,

042-Kudashev_I:_Layout 1 18.05.2008 15:35 Page 275

однако в словарных статьях рекомендуемые термины исходного языка и соответствия ставились первыми, а те единицы, которые наши специалисты-предметники считали нерекомендуемыми, снабжались знаком «недовольной мордочки» (☹). Например:

(1)
en **evaluation**
en **assessment**
fi **arviointi**
fi **evaluointi** ☹
ru **оценка**
ru **эвалюация** ☹

Хотя при простановке этой пометы мы в первую очередь ориентировались на мнение специалистов-предметников, окончательное решение принималось коллегиально, с опорой на авторитетные языковые ресурсы и с учетом мнений терминологов, редакторов и корректоров. Подобный подход позволяет более объективно фиксировать господствующую в том или ином подъязыке норму и в то же время учитывать мнение филологов. Очень важно, чтобы сведения о нерекомендуемости тех или иных терминов выражались в словаре эксплицитно, а не путем опущения нерекомендуемой лексики. Пользователь получает в свое распоряжение «информацию к размышлению», которая, безусловно, важна для него на этапе выбора варианта перевода. Следующим шагом в этом направлении могло бы стать эксплицитное указание причин, по которым специальная единица признана нерекомендуемой.

В общеязыковой лексикографии составление серьезного словаря уже практически немыслимо без электронного корпуса текстов. В терминографии процесс перехода к корпусным методам сдерживается рядом факторов. Во-первых, создание достаточного по объему корпуса и инструментов работы с ним – это серьезный проект, требующий инвестиций, времени и специальных знаний. Ограниченность рынка сбыта словарей специальной лексики ставит экономическую целесообразность такого проекта под сомнение (ср. Bergenholtz & Tarp 1995: 94). Другим фактором, сдерживающим использование в терминографии корпусов, является скорость устаревания научно-технической информации. За то время, которое необходимо для формирования корпуса, информация в нем может уже устареть. Наконец, к специальным текстам не всегда бывает просто получить доступ. Многие из них предназначены лишь «для внутреннего пользования». Будем надеяться, что в ближайшем будущем хотя бы часть этих проблем удастся преодолеть, и методы корпусной лингвистики и лексикографии получат более широкое распространение при составлении словарей специальной лексики.

Наличие определений и пояснений

Среди словарей, составляемых за пределами терминологических центров, редко встречаются словари с определениями и пояснениями. Еще более редкими являются словари, в которых определения составлены системно, на основе понятийного анализа. Основная причина этого заключается, по-видимому, в том, что составление определений является наиболее трудоемкой частью терминологического проекта. По оценкам специалистов финского Центра терминологической работы (Sanastokeskus TSK), на долю этапа понятийного анализа и составления определений приходится почти половина от общего времени работы над терминологическим словарем (Nykänen 1999: 66). Соответственно, при наличии определений сокращается количество терминов, которое авторский коллектив в состоянии обработать за отведенный промежуток времени, а также возрастает стоимость каждой словарной статьи. Кроме того, понятийный анализ и составление терминологически правильных определений требует соответствующей квалификации (не только предметной, но и терминологической). При этом конечные пользователи словарей, а вслед за ними и издатели часто предпочитают количество терминов качеству их разработки.

Тем не менее, включение определений в терминологические словари желательно по целому ряду причин. Во-первых, без четких определений трудно проводить сопоставление объема понятий в разных языках. Во-вторых, определение является источником ценной информации о специальных понятиях для переводчиков, которые, как правило, не являются специалистами-предметниками. В-третьих, определение позволяет переводчику сравнить значение термина, который приводится в словаре, со значением термина, который ему встретился в тексте. Это крайне важно, поскольку значение терминов может варьироваться от одного текста к другому, а также в диахронии.

При оценке необходимости определения понятий необходимо исходить из потребностей адресной группы словаря, а также учитывать следующие обстоятельства. Во-первых, для ряда специальных единиц, например, номенов, обозначающих объекты флоры и фауны, типы биоценозов и т. п., терминологическое определение составить невозможно, поскольку выделение соответствующих понятий производится не на основе родовидовых

042-Kudashev_I:_Layout 1 18.05.2008 15:35 Page 276

признаков. В этом случае определение заменяется пояснением, в котором перечисляются характерные признаки номена. Во-вторых, словник терминологического словаря не всегда можно представить в виде стройной системы понятий. В-третьих, не все единицы, входящие в словник, нуждаются в определении. Например, в «Словаре лексики проектного сотрудничества ЕС – Россия» определяются лишь специфические для данной предметной подобласти и трудные для понимания понятия. Лексика, привлеченная из других областей (например, экологическая, общеполитическая), приводится без определений.

Таким образом, в отношении определений описываемые словари отличаются от большинства других самим фактом наличия определений, их терминологическим характером (системность, соответствие логико-понятийным схемам), а также «ресурсосберегающей» методикой работы, заключающейся в том, что определения приводятся выборочно и обрабатываются терминологами до поступления специалистам-предметникам. Это позволяет создавать толково-переводные словари значительно большего, по сравнению со стандартами на термины и определения, объема.

Наличие логико-понятийных схем

Логико-понятийные схемы оказывают большую помощь при составлении терминологических определений, а также позволяют пользователям словаря получить наглядное представление о структуре предметной подобласти и отношениях между понятиями. В России логико-понятийные схемы в словарях практически не встречаются, а редкие исключения обычно оформляются в виде общего «понятийного поля», без графического выделения различных типов отношений между понятиями. В Финляндии включение логико-понятийных схем в словари фактически является нормой, причем графически выделяются три наиболее важных типа отношений – родовидовые, партитивные и функциональные.

Основные правила оформления логико-понятийных схем описаны в стандарте ИСО (ISO 704:2000: 6–14). Кроме того, финским Центром терминологической работы разработан ряд дополнительных способов графического представления логико-понятийных схем. При работе над «Финско-русским лесным словарем» использовались как стандартные способы оформления логико-понятийных схем, так и некоторые дополнительные приемы. Приведем несколько примеров логико-понятийных схем:

20.27 Устройство осушительной сети

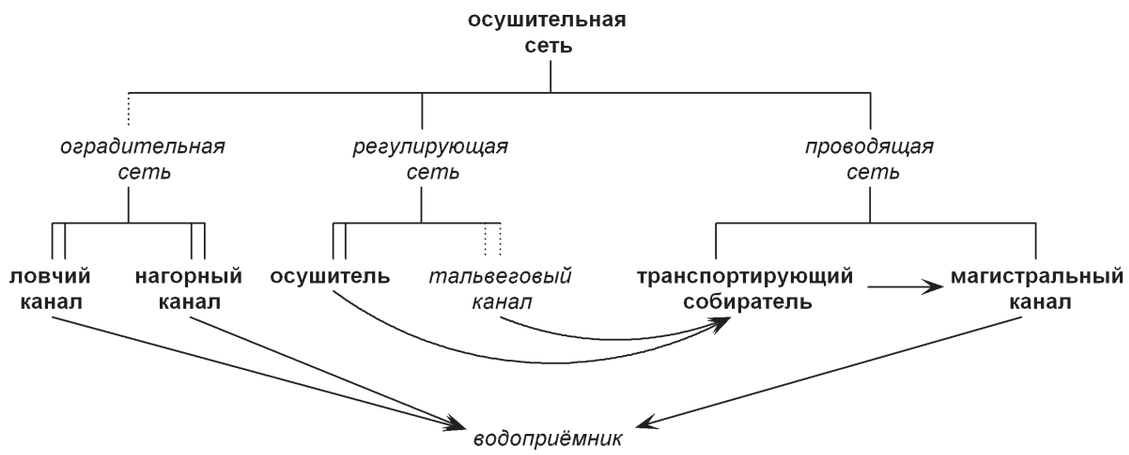


Рис. 1

В приведенных схемах расходящиеся линии отображают родовидовые отношения, «гребневидные» – отношения «часть-целое», а стрелки – функциональные (неиерархические) отношения. Характер функциональных отношений иногда поясняется вербально, как это сделано в схеме 20.20. Двойные линии свидетельствуют о множественности объектов (например, регулирующая сеть всегда содержит несколько осушителей). Пунктир обозначает факультативность класса (например, осушительная сеть может содержать оградительную сеть). Термины, выделенные курсивом, важны для целостного восприятия схемы, но не вошли в корпус словаря.

042-Kudashev_I:_Layout 1 18.05.2008 15:35 Page 277

20.28 Деформация канала

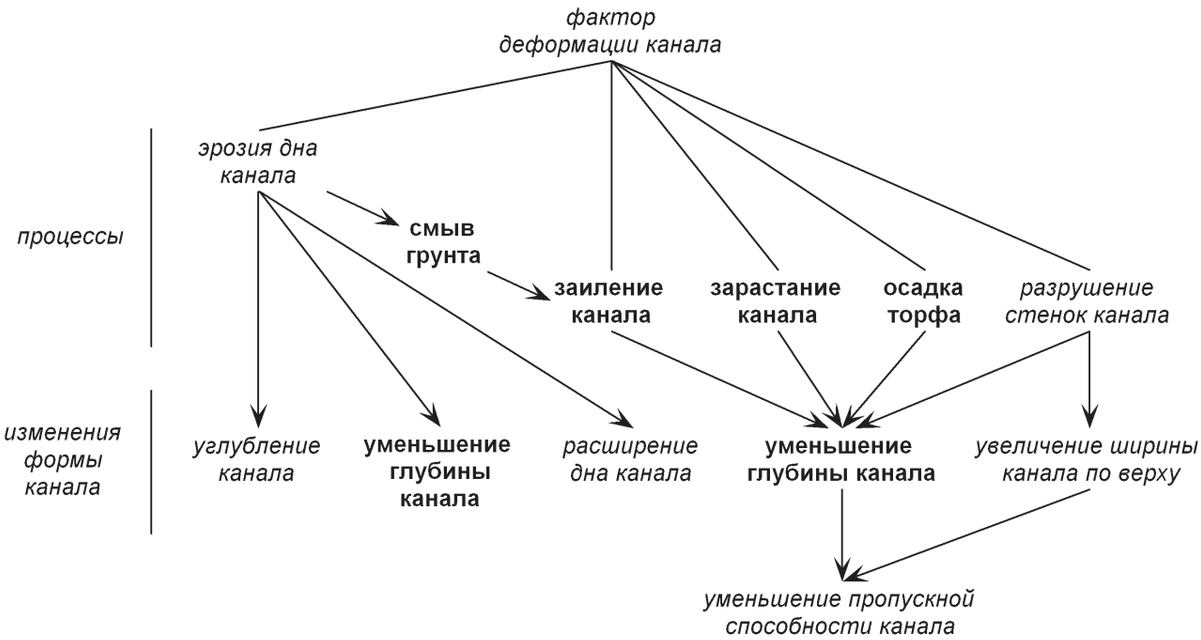


Рис. 2

20.20 Гидрологические последствия гидролесомелиорации на осушенной территории

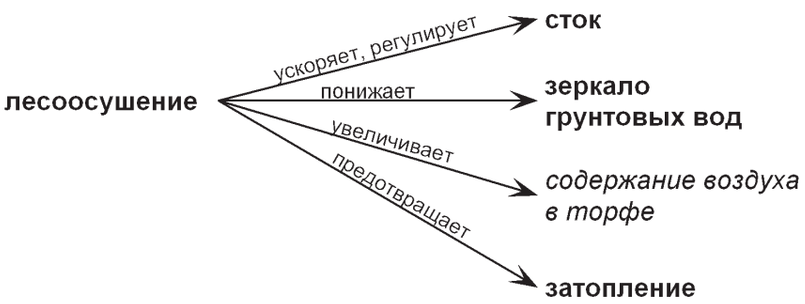


Рис. 3

В приложении к «Финско-русскому лесному словарю» приводятся свыше 700 подобных логико-понятийных схем на финском и русском языках. Схемы составлялись отдельно для финской и русской системы понятий, т. е. они не являются переводом друг друга. Из корпуса словаря к схемам сделаны отсылки.

Наличие указаний на источники

В отличие от стандартов на термины и определения, которые утверждаются комиссиями специалистов, в описательных словарях желательны указания на источники сведений. Они служат своего рода «индексом надежности» и «индексом свежести» информации, а также сообщают пользователю о том, где он может получить дополнительную информацию о понятии. В случае переводных соответствий указание на источник позволяет также определить, является ли переводное соответствие реально использующимся или искусственно созданным.

042-Kudashev_I:_Layout 1 18.05.2008 15:35 Page 278

В «Финско-русском лесном словаре» приводятся ссылки на источники всех терминов, определений, пояснений и переводных соответствий. Письменные источники помещаются за знаком раскрытой книги. Если определения или пояснения основываются на каком-либо источнике, но подверглись модификации, после знака книги ставится знак «приблизительно равно». Например:

(2)
karike [Metsäekologian perusteet] [Metsämaa]
puustosta ja muusta kasvillisuudesta maanpinnalle variseva tuore tai heikosti hajonnut eloperäinen aines 🕮 ≈ SESMS
ⓘ Maahan tulee eloperäistä ainesta myös juurikarikkeena. 🕮 ≈ Mälkönen 2003, 82
расти́тельный опа́д 🕮 ОСТ 56-108-98
опа́д 🕮 Белов, 128

Термины и соответствия, по каким-либо причинам не найденные в письменных источниках, но подтвержденные специалистами-предметниками, снабжаются галочкой и фамилией специалиста. Такая практика считается менее предпочтительной по сравнению с указанием на письменный источник, однако в ряде случаев она позволяет сэкономить немало времени. Определения и пояснения, составленные специалистом, помечаются знаком ручки и фамилией специалиста. Например:

(3)
haaraisuus (23.17) 🕮 Kärkkäinen 2003, 250 [Puutavaralajit ja laatuvaatimukset]
vikaisuus, joka ilmenee puujoukossa päärangasta erottuvina haaroina ✍ Kärkkäinen
разветвлённость ствола́ *f* ☑ Филипчук

В приложении к словарю приводится полный список письменных источников, а также список специалистов-предметников с указанием научных степеней, званий, места работы и должности.

Указание на характер и степень соответствия

Одним из широко распространенных недостатков переводных терминологических словарей является подача вариантов перевода через запятую или точку с запятой безо всяких комментариев относительно характера и степени соответствия. В «Энциклопедии лексикографии» такая практика названа одним из «смертных грехов переводной лексикографии» (Kromann, Riiber & Rosbach 1991: 2724).

В описываемых словарях частичные соответствия снабжены знаком ≈ и пояснениями о сути различий между понятиями. Например:

(4)
metsämaa 🕮 SESMS [Kasvupaikka- ja metsätyypit] [Metsätalouden suunnittelu] [Puun ja metsän mittaus]
metsätalousmaa, jonka puuntuotoskyky on vähintään 1 m³/ha ✍ Poso
≈**покры́тые лесно́й расти́тельностью зе́мли** *pl.* 🕮 ОСТ 56-108-98
≈**лесопокры́тые зе́мли** *pl.* 🕮 ОСТ 56-108-98
Финский термин обозначает лесохозяйственные угодья, продуктивность которых составляет не менее 1 м³/га. В России к землям, покрытым лесной растительностью, относятся насаждения с полнотой 0,3 и выше и запасом не менее 50 м³/га. ✍ Филипчук

metsäpaloalue ☑ Vanha-Majamaa [Metsänhoito] [Metsäpalot]
kuloalue 🕮 Berninger et al., 139
paloalue 🕮 SESMS
kuloala 🕮 SESMS
metsäpalossa palava tai palanut alue ✍ Vanha-Majamaa
≈**террито́рия лесно́го пожа́ра** ☑ Сергеева
Финский термин может обозначать как территорию, на которой действует лесной пожар, так и территорию, пройденную лесным пожаром. Термин «территория лесного пожара» обозначает территорию, на которой действует пожар. ✍
≈**гарь** *f* 🕮 ЭЛХ I
Территория, пройденная пожаром, с полностью уничтоженным древостоем. ✍ Сергеева
≈**горе́льник** 🕮 ЭЛХ I
Территория, пройденная пожаром, с остатками древостоя. ✍ Сергеева

042-Kudashev_I:_Layout 1 18.05.2008 15:35 Page 279

en **dissemination of experience**
fi **kokemusten levittäminen**
def. kokemuksista tiedottaminen ✎
def. информирование о каком-либо опыте
ru **распространение опыта**
ru ≈**передача опыта**
Распространение опыта подразумевает только распространение информации об опыте, в то время как передача опыта обычно подразумевает также практические шаги (например, обучение).

После искусственно созданных соответствий ставится знак ручки, поскольку переводчик должен знать об искусственном происхождении эквивалента, чтобы правильно выбрать способ его введения в текст. Дело в том, что в тексте перевода искусственные эквиваленты часто приходится снабжать, например, кавычками или дополнительными пояснениями. Если соответствие предложено специалистом-предметником, а не авторским коллективом словаря, после ручки ставится фамилия специалиста. Например:

(5)
estimaattori 🕮 Kangas et al., 8 [Puun ja metsän mittaus]
matemaattinen malli, jolla arvioitavalle suureelle hankitaan arvio ✎ Poso
ⓘ Yleensä kyseessä on otantamenetelmään liittyvä tilastotieteellinen malli. ✎ Poso
моде́ль статисти́ческого ме́тода *f* ✎ Филипчук

Наличие глаголов и терминоэлементов

Глаголы играют в языках для специальных целей важную роль, однако в терминологических словарях они часто подвергаются дискриминации в пользу субстантивной и субстантивированной лексики. При этом переводчики лишаются информации о форме глагола, ее видоизменениях, моделях управления и т. д. Мы посчитали необходимым включить эту информацию в наши словари. Приведем примеры глагольных словарных статей:

(6)
karsia [Puunkorjuu]
poistaa puusta oksia tai oksat 🕮 CD-PS
очища́ть от су́чьев (*perf.* очистить) 🕮 ЛГ
◊ karsia puu, karsia oksia puusta – *очищать дерево от сучьев*

katkoa 🕮 CD-PS [Puunkorjuu]
pölkyttää ☑ Kärkkäinen
katkaista runko useaan osaan 🕮 ≈ CD-PS
осуществля́ть раскряжёвку (*perf.* осуществить) ☑ Филипчук
выполня́ть раскряжёвку (*perf.* выполнить) ☑ Филипчук
раскряжёвывать (*perf.* раскряжевать) 🕮 АРРА
◊ katkoa runko – *осуществлять* (*выполнять*) *раскряжёвку ствола*, *раскряжёвывать ствол*

В «Словаре проектной лексики ЕС – Россия» наиболее характерные сочетания с глаголами приводятся также в субстантивных статьях, например:

(7)
en **tender competition**
en **call for tenders**
en **call for bids**
fi **tarjouskilpailu**
def. menettely, jossa pyydetään tarjouksia usealta tuotteen tai palvelun toimittajalta 🕮 ≈ CD-PS
def. процедура, в ходе которой поставщиков товаров или услуг просят представить тендерные заявки
ru **тендер**
ru **конкурс предложений**
ru **конкурсный отбор** ☹
◊ to conduct a tender competition – järjestää tarjouskilpailu (kilpailuttaa) – *проводить тендер*; to win a tender competition – voittaa tarjouskilpailu – *выигрывать тендер*; public procurements should be put out to tender – julkiset hankinnat on kilpailutettava – *на публичные закупки должны проводиться тендеры*

042-Kudashev_I:_Layout 1 18.05.2008 15:35 Page 280

В «Финско-русский лесной словарь» наряду с глаголами включались также некоторые продуктивные терминоэлементы. Представление терминоэлементов помогает сэкономить место в словаре путем демонстрации типичных моделей терминообразования в исходном языке и моделей образования переводных эквивалентов. Например, с финским прилагательным pyöräalustainen (колёсный, на колёсном ходу) может сочетаться большое число существительных, причем перечисление их всех в словаре невозможно и нецелесообразно. Если же в статье pyöräalustainen дать несколько типичных примеров сочетаний с данным терминоэлементом, то это послужит образцом для перевода многих других аналогичных сочетаний:

(8)
pyöräalustainen 🕮 Uusitalo, 60 [Koneet ja laitteet]
pyörillä varustetulla alustakoneella sijaitseva 🖉 Kärkkäinen
колёсный 🕮 Ууситало, 62
на колёсном ходу́ 🕮 Жаденов & Заикин, 61
◊ pyöräalustainen kaivinkone – *колёсный экскаватор*, *экскаватор на колёсном ходу*; pyöräalustainen metsäkone – *лесохозяйственная* (*лесосечная*) *машина на колёсном ходу*

Заключение

Нельзя сказать, что приемы, описанные выше, не были известны ранее в практике мировой лексикографии. Однако подавляющая масса неодноязычных терминологических словарей, издаваемых в настоящее время, по-прежнему строится по принципу «термин-перевод» и практически не содержит дополнительных сведений, позволяющих переводчику правильно выбрать иноязычное соответствие и быть уверенным в своем выборе. В этом смысле описанные в статье словари являются, на наш взгляд, большим шагом вперед. Отдельные приемы, использованные в них, могут модифицироваться и дорабатываться в соответствии с требованиями конкретных словарных проектов. Например, наряду со знаками для обозначения письменных и устных источников может потребоваться введение специального знака для Интернет-источников. Типы функциональных отношений в логико-понятийных схемах могут конкретизироваться вербальными или графическими средствами. Даже сами типы отношений могут быть пересмотрены и приведены, например, в соответствие с теми, которые используются при построении онтологий. Однако общие принципы описанных словарей – приведение определений, пояснений, логико-понятийных схем, указание на источники, маркирование частичных и искусственно созданных эквивалентов, обязательное наличие комментариев о различиях между разноязычными понятиями и терминами – эти принципы, как нам представляется, должны в перспективе стать нормой при разработке качественных переводных терминологических словарей.

Список литературы

1. ISO 704:2000. Terminology Work – Principles and Methods. Geneva: ISO.

2. ISO 12620:1999. Computer Applications in Terminology – Data Categories. Geneva: ISO.

3. Татаринов В.А. Теория терминоведения: В 3 т. – Т. 1. Теория термина: История и современное состояние.– М.: Моск. Лицей, 1996. – 311 с.

4. Bergenholtz, H. & Tarp, S. (eds) Manual of Specialised Lexicography. The Preparation of Specialised Dictionaries. Amsterdan: Benjamins, 1995.

5. Nykänen, Olli. Sanastoprojektin vaiheet // Toimikunnista termitalkoisiin. 25 vuotta sanastotyön asiantuntemusta. Helsinki: Tekniikan Sanastokeskus, 1999. 62–71.

6. Kromann, H.-P., Riiber, T. & Rosbach, P. Principles of Bilingual Lexicography // Hausmann, F.J. et al. (eds) Wörterbücher: ein internationales Handbuch zur Lexikographie. Berlin: de Gruyter, 1991. T. 3. 2711–2728.

043-Kuznetcov_I:_Layout 1 18.05.2008 15:37 Page 281

ОСОБЕННОСТИ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ ТЕКСТОВ СЕМАНТИКО-ОРИЕНИРОВАННЫМ ЛИНГВИСТИЧЕСКИМ ПРОЦЕССОРОМ SEMANTIX

LINGUISTIC PROCESSOR "SEMANTIX" FOR KNOWLEDGE EXTRACTION FROM NATURAL TEXTS IN RUSSIA AND ENGLISH

Кузнецов И.П. *(igor-kuz@mtu-net.ru), ИПИ РАН,*
Ефимов Д.А. *(d.efimov@synsys.ru), ЗАО Синергетические Системы.*

Лингвистический процессор Semantix предназначен для областей, где требуется автоматическая формализация потоков текстов на естественном языке: резюме, сообщения СМИ, информационно-рекламные материалы, почтовые сообщения, сводки происшествий, справки по уголовным делам, архивные материалы и др. Из текстов (документов) извлекаются интересующие пользователя объекты, их свойства и связи. Представляются факты участия объектов в действиях. Последние сами рассматриваются как комплексные объекты с их свойствами и связями. В результате на основе каждого документа строится специального вида семантическая сеть, отражающая его семантическую структуру. Такие сети отображаются на XML-файлы, которые служат для организации Баз Знаний, соответствующих семантических поисков, для решения логико-аналитических задач, а также для автоматического заполнения реляционнных БД.

Введение

Исследования ведущих аналитиков показывает, что совокупный объем цифровой информации в 2006 году составил 161 миллион гигабайт. Предполагается, что за период с 2006 по 2010 год объем информации увеличится более чем в шесть раз. В более чем 80% случаев такая информация является неструктурированной - это тексты естественного языка. Человеку становится все труднее ориентироваться в потоках поступающей информации. В связи с этим при обработке информации требуются новые инновационные подходы, ориентированные на задачи конкретных пользователей.

Следует учитывать, что большая категория пользователей имеют определенные служебные обязанности, и соответственно, постоянные интересы. Им необходима вполне конкретная информация. Например, сотрудники информационно-аналитических подразделений выбирают из СМИ информацию об интересующих их событиях, катастрофах, террористических актах, персоналиях и др. Следователю важны фигуранты, места их жительства, телефоны, криминальные события, даты и др. Сотруднику кадровой службы нужно знать организации, где, кем и в какое время кандидат работал. Подобная информация называется ***информационными объектами*** или просто ***объектами*** [1,2,10]. Объекты различаются по ***типам***. Каждая из перечисленных категорий пользователей интересуется набором объектов определенного типа. Находить нужные объекты в потоке текстов, читая их, во многих областях - непосильный труд.

Для обеспечения подобных пользователей нужной информацией требуются средства автоматического извлечения объектов из текстов с их представлением в формах, удобных для восприятия или последующей обработки. Речь идет об автоматической формализации текстов, связанной с ***извлечением знаний*** (Knowledge Extraction). Это проблемная область, которая находится в сфере внимания исследователей. Ее актуальность постоянно растет [3,4,5]. Особенность наших исследований – в их ориентации на логико-аналитическую обработку. Для этой цели на протяжении последних 15 лет в рамках проектов ИПИ РАН разрабатывались семантико-ориентированные лингвистические процессоры для аналитических служб. Первый процессор построен более 10 лет назад для логико-ангалитической системы Криминал [6,7]. Их научная база: ***расширенные семантические сети (РСС)***, методики представления сложных видов знаний, ***инструментальная среда ДЕКЛ*** обработки структур знаний, сетевые позиционные грамматики, онтологии в формате РСС, морфологический анализ на основе обобщенных окончаний [1,2,14]. Последний вариант такого процессора, изготовленного совместно с ЗАО <Синергетические Системы> в виде модуля SDK, получил название ***Semantix***.

043-Kuznetcov_I:_Layout 1 18.05.2008 15:37 Page 282

1. Основные компоненты процессора Semantix

Лингвистический процессор Semantix предназначен для областей, где требуется автоматическая обработка потоков текстов на ***естественном языке (ЕЯ***): резюме, сообщения СМИ, информационно-рекламные материалы, почтовые сообщения, сводки происшествий, справки по уголовным делам, архивные материалы и др. Из текстов (документов) выделяются интересующие пользователя объекты, их связи, а также факты участия объектов в тех или иных действиях или событиях. Последние сами рассматриваются как комплексные объекты с их свойствами и связями. В результате на основе каждого документа строится специального вида семантическая сеть (РСС), представляющая его ***семантическую структуру***. Такая сеть отображается на XML-файл. С их помощью значительно облегчается последующий автоматический анализ. XML-файлы являются основой для составления досье, обзоров, отчетов. Другой вариант их использования - автоматическое заполнение реляционных БД или формирование собственной ***Базы Знаний*** с последующей организацией направленного поиска нужной информации (объектов), в том числе, различных видов семантического поиска.

Основные компоненты процессора Semantix:

1.1. Блок лексического и морфологического анализа. Выделяет из текста слова и предложения, приводит слова нормальную форму и формирует семантическую сеть, представляющую ***пространственную структуру текста (ПС***), где отображается последовательность слов, их основные признаки, начало предложений и наличие пробельных строк. Блок использует специальный набор тематических словарей (словарь стран, регионов России, имен, видов оружия и др.) для группирования слов и придания им дополнительных семантических признаков [14].

1.2. Блок синтактико-семантического анализа. Он преобразует одну семантическую сеть (ПС) в другую, представляющую ***семантическую структуру текста (СС***), т.е. выделенные объекты и их связи. Последнюю часто называют ***содержательным портретом документа*** [9,10]. Блок управляется ***лингвистическими знаниями (ЛЗ),*** которые определяют процесс анализа текста. ЛЗ включают в себя специального вида контекстные правила, которые обеспечивают высокую степень избирательности при выявлении (извлечении) объектов и связей [8].

Задачи этого блока:

- Извлечение из потока ЕЯ-документов информационных объектов: лиц, организаций, действий, их места и времени, и многих других объектов.

- Выявление связей объектов. Например, как лица связаны с организациями (*МЕСТО_РАБОТЫ*), адресами (*ПРОЖИВАЕТ, ПРОПИСАН*). Или как фигуранты связаны с объектами типа оружие, наркотики (ИМЕТЬ).

- Анализ глагольных форм, причастных и деепричастных оборотов с выявлением фактов участия объектов в соответствующих действиях. Например, один фигурант передал другому фигуранту наркотики – это факт, связывающий фигурантов.

- Выявление связей действий с объектами типа место или время (где и когда имело место данное действие или событие).

- Анализ причино-следственных и временных связей между действиями и событиями.

1.3. Экспертные системы (ЭС). На основе сети СС формируют новые знания - в виде дополнительных фрагментов РСС. Например, при обработке тектов резюме по каждой автобиографии ЭС выявляют область деятельности лица по его автобиографии (в соответствии с заданным классификатором). Оценивается опыт его работы. При анализе криминальных действий ЭС осуществляют соотнесение криминального происшествия к определенному типу: выявляют характер преступления, способ его совершения, орудие и т.д. (в соответствии с классификаторами криминальной милиции).

1.4. **Обратный лингвистический процессор**, преобразующий содержательный портрет документа (семантическую сеть СС) в XML-файл. При этом осуществляются необходимые замены символов, служебных слов (имен объектов), выставляются метки начала и конца объектов, действий, предложений. Преобразование осуществляется без потери информации. XML-файл устроен таким образом, что в нем представлены все выявленные компоненты и связи. В случае необходимости, обеспечивается обратное преобразование XML-файл в сеть СС.

1.5. База лингвистических и экспертных знаний (БЗ). Содержит правила анализа текста и экспертных решений во внутреннем представлении. Они определяют работу лингвистического процессора. Semantix имеет несколько таких баз, которые активизируются в зависимости от предметной области и задач пользователя, см. п.4.

043-Kuznetcov_I:_Layout 1 18.05.2008 15:37 Page 283

2. Выделяемые объекты и связи

Набор выделяемых объектов зависит от задач пользователя. В тоже время, качество лингвистического процесора в значительной степени определяется возможностями такого выделения. Ниже перечислены основные ***типы информационных объектов*** и связей, извлекаемые Semantix:

- лица (по ФИО) с их особенностями (потерпевший, террорист и др.);
- адреса, почтовые атрибуты;
- организации;
- должности;
- террористические группы, ОПГ;
- номера телефонов, факсов, электронных постовых адресов с их стандартизацией;
- средства транспорта с выделением марки машины, государственного номера, цвета и других атрибутов;
- количественные характеристики (сколько лиц или других объектов принимали участие в том или ином событии);
- паспортные данные и другие документы с их атрибутами;
- взрывчатые вещества;
- наркотические вещества;
- оружие с атрибутами;
- словесное описание лиц, их приметы;
- номера счетов, суммы денег с указанием типа валюты;
- события (криминальные, террористические, поломки изделий и др.) с указанием участия в них информационных объектов;
- время и место событий;
- связи между различными типами информационных объектов, включая комплексные объекты (действия или события);
- другие объекты (опыт работы, знание языков ... до 40 типов).

На рис.1 представлено графическое изображение этих объектов в ***ДЕМО-версии***. (ДЕМО-версия в сети Интернет находится на сайте www.semantix4you.com).

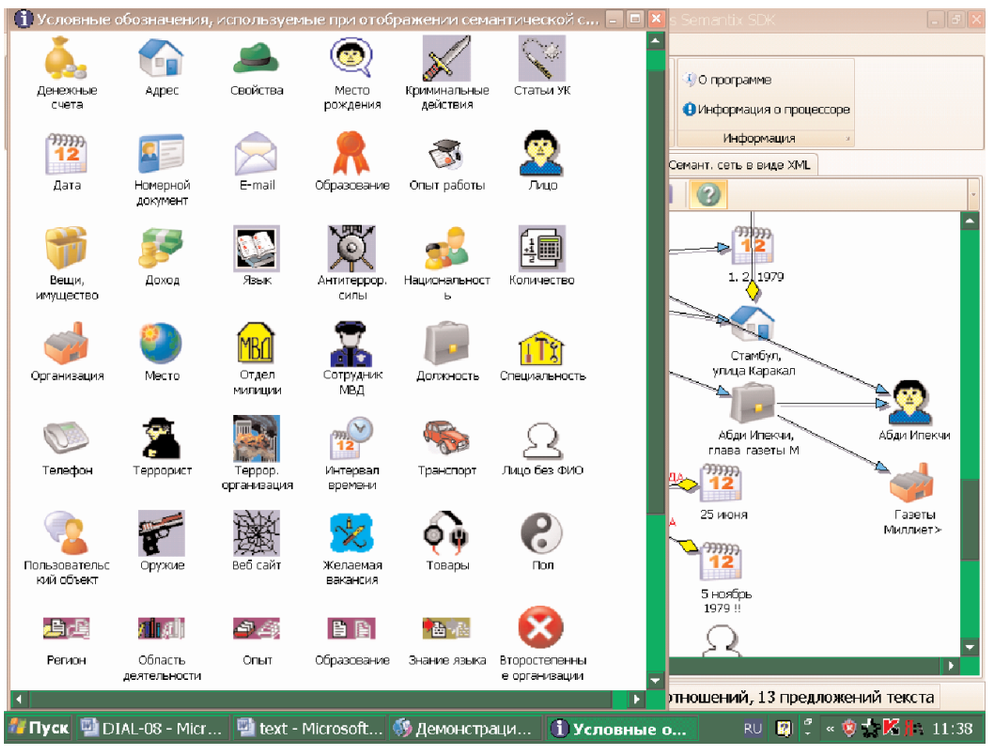


Рис.1. Набор выделяемых объектов процессором Semantix

043-Kuznetcov_I:_Layout 1 18.05.2008 15:37 Page 284

При выделении объектов учитываются возможные варианты называния объекта в тексте, в том числе, в краткой форме. Типовые объекты (ФИО, даты, адреса, должности и др.) приводятся к одному (стандартному) виду. Осуществляется идентификация объектов с учетом кратких наименований (например, отдельных фамилий или имен с ФИО), анафорических ссылок (указательных и личных местоимений, например, «*Этот человек», «Он ...»*), определений (например, *«Мэр Москвы Лужков»* идентифицируется с последующими словами «*мэр», «Лужков*«).

Выделение связей - это не только глубинный анализ глагольных и других форм. Многие связи даются по умолчанию. Например, в сводках происшествий, как правило, за ФИО фигуранта следуют его данные без указания их принадлежности и с дополнительными текстовыми вставками. В связи с этим в процессоре Semantix для ряда объектов организуется направленный поиск связанных объектов, т.е. восстановление связей, данных по умолчанию. Для этого организуются специальные процессы, чтобы связать лицо с его местом проживания или местом работы, принадлежащим ему автотранспортом и т.д. Наример, при анализе сводок происшествий это делается следующим образом. Для ряда объектов (адрес, телефон, г.рожд и др.) строится виртуальная связь с другими объектами (ФИО, организации), пока неизвестными. Далее, на одном из уровней обработки с помощью специальных ***правил идентификации*** производится их поиск. В этих правилах указывается направление поиска, допустимое количество шагов, а также признаки слов и знаки препинания, где процесс поиска следует закончить. При этом требуются специальные фильтры, чтобы не захватить и не связать посторонний объект. Такой подход показал достаточно хорошие результаты в системе Криминал [6].

В результате строится РСС, называемая ***содержательным портретом документа***. При этом учитываются особенности ЕЯ, где с помощью глаголов, отглагольных существительных и причастных оборотов задаются одни и те же действия. При представлении на РСС они приводятся к одному виду – комплексному объекту. Более того, формы с отглагольными существительными могут быть компонентами глагольных форм. Аналогично, в РСС одни объекты могут быть компонентами других. Представляются причино-следственные и временные зависимости между действиями, событиями, которые отражают логическую связь предложений, заданную в явном виде – с помощью слов типа *поэтому, затем* и др. Пример содержательного портрета, изображенного в виде графа, представлен на Рис.2.

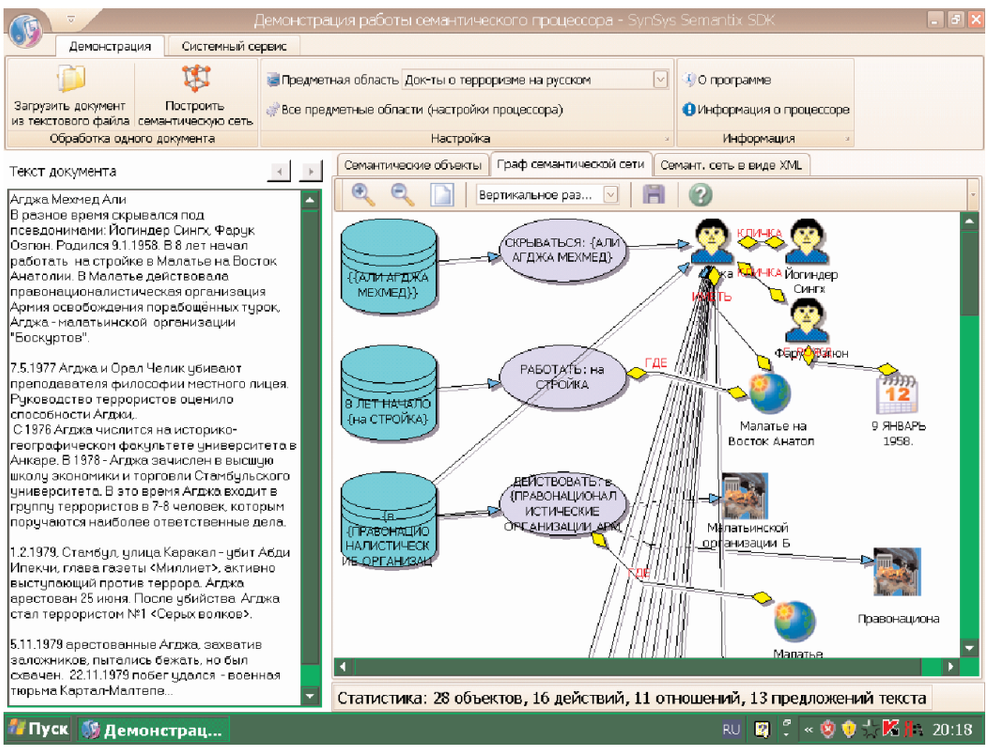


Рис.2. Графическое представление содержательного портрета документа

На данном примере видно, что фигурант *Агджа Мехмет Али* во многих случаях задается его именем *Агджа* и в результате идентификации имеет много связей. С помощью эллипсов изображаются действия, которые связываются с предложениями.

043-Kuznetcov_I:_Layout 1 18.05.2008 15:37 Page 285

3. Факторы, определяющие качество процессора

Качество лингвистического процессора определяется рядом факторов. Во-первых, это возможности выделения объектов и связей. Имеется в виду типы выделяемых объектов, их количество. Процессор Semantix выделяет до 40 типов объектов, в том числе комлексных объектов, соответствующих действиям и событиям. С увеличением количества возникают дополнительные трудности, связанные с «коллизией» правил выделения: одни правила могут захватывать слова, относящиеся к другим объектам и выделяемым другими правилами. становится важным порядок применения правил, в том числе, правил идентификации..

Во-вторых, важный фактор - это ***избирательность правил*** и процедур идентификации**:** коэффициент шумов и потерь. Под ***шумами*** понимается наличие лишних слов в объектах. ***Потери*** - это когда объект не выявлен или выявлен частично: в тексте есть слова, которые не вошли в объект. В процессоре Semantix правила устроены таким образом, что они обеспечивают высокую степень избирательности и минимизацию шумов и потерь при большом количестве выделяемых объектов, см. п.3.

Третий фактор - возможность и трудоемкость ***настройки*** на корпус текстов (для повышения избирательности правил выделения объектов), а также настройки на новые объекты. В связи со сложностью процессов анализа такая настройка должна осуществляться через ***лингвистические знания*** (ЛЗ). Последние должны иметь все средства для повышения избирательности правил и необходимые удобства в плане их создания и корректировки. В идеале с помощью ЛЗ должна обеспечиваться настройка на особенности языка - признаки, которые даются словам, на типовые конструкции и формы языка. Лингвистический процессор должен быть в значительной степени индифирентен к языку. Его задача - поддерживать ЛЗ, в том числе, процесс применения правил выделения идентификации.

По такому принципу организован процессор Semantix, в котором за счет ЛЗ обеспечивается анализ сложных конструкций русского языка, а также анализ англо-язычных конструкций и форм, выделение англо-язычных объектов и их связей. Другими словами, обеспечивается анализ не только русского, но и английского языка. Это говорит об универсальности процессора.

Четвертый фактор - скорость работы лингвистического процессора, т.е. ***время анализа*** текстов. Скорость определяется конструктивными особенностями процессора (средствами уменьшения переборов), а также количеством выделяемых объектов. Применение правил их выделения связано с поиском нужных слов, где требуются переборы. Чем больше объектов и правил, тем больше переборов и больше время анализа.

В процессоре Semantix имеются различные средства уменьшения переборов. Помимо программных, также имеются средства, управляемые с помощью ЛЗ. Для каждого правила указывается, какие слова следует искать для инициирования процесса его применения. Задаются допустимые контексты (слева и справа от выявляемых слов), факультативные элементы [8].Таким образом обеспечивается достаточно высокая скорость (доли секунды на 1 кб. текста) при достаточно большом количестве выделяемых объектов. Отметим, что если объектов мало, то скорость значительно возрастает. В связи с этим в ЛЗ введены специальные средства, использующие список значимых слов и признаков (указывающих на наличие объектов) для выделения значимых предложений. Только их следует анализировать. И если в тексте много предложений без объектов, то таким образом скорость можно увеличить на порядки.

Следует отметить, что в связи с актульностью, область *извлечение знаний* (выявление объектов и связей) привлекает все больше исследователей, которые строят свои лингвистические процессоры. Хотя у них много общего. Используются правила синтактико-семантического анализа, которые по контексту выявляют объекты. Такие правила называют также шаблонами, фреймами, а их компоненты – элементами, узлами, слотами и др. Отличие - в используемых формализмах и наличии специальных (интеллектуальных) средств повышения избирательности правил, устранения коллизий. Например, лингвистический процессор системы Арион строился аспирантом в рамках коллектива ИПИ РАН, уже имеющего отлаженный процессор системы Криминал. Были заимствованы многие методики и подходы, которые в системе Арион реализованы в рамках формализма XML. Но не все. Поэтому (сколь известно автору - руководителю аспиранта), данный процессор пока не сумел достичь качества оригинала.

Процессор компании «Эр Си О», по-видимому, имеет достаточно хорошие средства установления анафорических связей [15]. Говорится о возможности извлечения фактов. Однако, примеры носят искусственный характер и остается непонятным степень работоспособности процессора на реальных текстах. Недавно появившийся процессор Ontos SOA, как говорится в рекламных материалах, содержит полный цикл сематико-синтаксической обработки. Имеется база знаний. Но ничего не говорится об их основе, которая обычно или заимствуется, или разрабатывается, что требует весьма продолжительного времени.

Отметим класс систем, в которых не требуется выделения разнообразных объектов и связей. Дело в том, что некоторые объекты (обычно это русские ФИО, даты, телефоны, организации в стандартной записи) доста-

043-Kuznetcov_I:_Layout 1 18.05.2008 15:37 Page 286

точно надежно выделяется и чисто программными средствами, без ЛЗ. Однако гибкость таких процессоров невелика.

4. Предметные области

Настройка на предметную область осуществляется при наличии соответствующего корпуса текстов путем разработки лингвистических знаний (ЛЗ), определяющих набор выделяемых объектов и связей. У коллектива разработчиков имеется большой опыт настройки на различные предметные области и корпуса текстов - для русского и английского языков, см. рис.3. Результатом являются отлаженные правила ЛЗ, обеспечивающие выделение большого количества разнотипных объектов (до 40 типов).

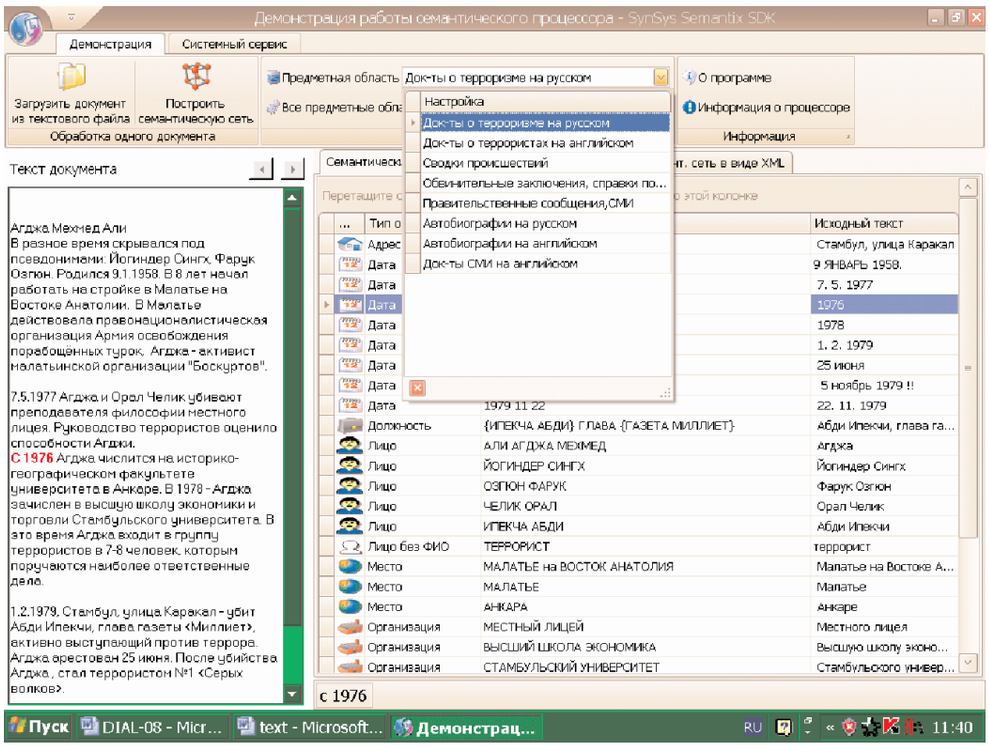


Рис.3. Предметные области, на которые настроен процессор Semantix

Рассмотрим эти области более подробно:

4.1. *Документы о терроризме на русском языке*. Анализ документов, в которых речь идет о террористических актах и группах. Обеспечиваетя выделение до 40 типов объектов, их связей и степень участия в криминальных действиях.

4.2. *Документы о террористах на английском языке*. Выделяются руководящие и другие лица, должности, организации, террористические группы, орудия преступления, время и место событий и т.д., а также связи и участие в действиях.

4.3. *Сводки происшествий*. Обеспечивается выделение фигурантов, их связей, организаций, дат, документов, номеров счетов, оружия ... (до 40 типов объектов) с указанием их участия в криминальных действиях

4.4. *Обвинительные заключения*, справки по уголовным делам. Объекты идентифицируются по всему полю текста. Выявляются их связи и криминальные действия.

4.5. *Правительственные сообщения*, СМИ. Выделяются лица, даты, организации, должности и другая значимая информация, а также связи и участие в действиях.

4.6. *Автобиографии на русском языке*. Из русскоязычных резюме выделяются все атрибуты человека, периоды времени и место его работы, учебы, знание языков и т.д.

4.7. *Автобиографии на английском языке*. Из англоязычных резюме выделяются все атрибуты человека (см.п. 4.6.).

043-Kuznetcov_I:_Layout 1 18.05.2008 15:37 Page 287

4.8. *Документы СМИ на английском*. Из англоязычных текстов выделяются упомянутые в СМИ лица, должности, организации, даты, террористические и антитеррористические группы, оружие, события, их время. место, различные связи и др.

Как результат достаточно эффективного процесса настройки на различные предметные области, в Semantix имеется достаточно большой набор правил избирательного выявления из текстов разнотипных объектов.

Первые правила, осуществляющие выделение дат, адресов, лиц, автотранспорта, криминальных объектов (оружие, наркотики) и др., отлаживались на корпусе текстов ГУВД г. Москвы: сводки происшествий, справки по уголовным делам, записные книжки фигурантов и др. (более 500 тыс. документов). Никаких ограничений на тексты не накладывалось. И этого нельзя было делать, так как большие потери криминальной информации недопустимы. При этом удалось добиться уникальных результатов. Коэффициент шумов удалось свести до уровня, не превышающего 1-2%, а коэффициент потерь около 1%.

Далее ЛЗ были настроены на выделение объектов из автобиографий, написанных на русском языке. При этом потребовалась настройка на значительное количество объектов нового типа [13]. Соответствующие правила отлаживались на корпусе текстов, состоящих более чем из 1000 резюме. Стояла задача обработки любых текстов резюме с возможностью использования процессора для компаний, работающих до настоящего времени с неформализованными резюме.

Далее, процессор был настроен на работу с резюме на английском языке. Использовался корпус текстов около 500 резюме. Построение англоязычного процессора на базе русскоязычного носило в большей степени экспериментальный характер. В процессор были добавлены средства, учитывающие особенности английского языка – словообразование, многозначность слов и др. При этом удалось добиться достаточно хорошего качества [16].

Следующий этап - это тексты СМИ с дополнительным выделением террористических организаций, групп, отдельных лиц, а также сил, противоборствующих терроризму. Потребовались дополнительные правила для выделения арабских ФИО, идентификации объектов и др. Правила ЛЗ отлаживались на корпусе текстов около 1000 сообщений СМИ, правительственных сообщений и материалов из других источников (документы от 2-х до 40 кб.). Далее за счет ЛЗ процессор был настроен на работу с документами СМИ на английском языке [12]. Результатом явилось большое количество отлаженных правил выделения объектов из различных текстов русского и английского языков. В рамках системы Semantix пользователю предоставляется возможность выбора этих объектов, см. Приложение. Еще раз отметим, что если пользователю не требуется анализа предложений или его не интересуют какие-либо объекты из заданного перечня, то он указывает это в соответствующем меню. В результате скорость анализа может возрасти на порядок.

5. Структура XML-файла

В XML-файле представлена содержательный портрет (структура СС), т.е. все объекты и связи, выявленные их текста процессором Semantix. В связи с этим организация XML-файлов имеет определенный научный интерес – как средств представления семантической структуры предложений и текстов.

Преобразование сети СС в XML-файл обеспечивается с помощью обратного лингвистического процессора. При этом фрагменты, представляющие объекты, отношения, действия и предложения в структуре СС, отображаются на соответствующие компоненты XML-файла, которые также будем называть объектами, отношениями, действиями и предложениями. Рассмотрим основные компоненты, из которых состоит XML-файл.

5.1. ***Константа*** - это простейшая компонента СС, представляющая собой одно нормализованное слово или символ ЕЯ. Под нормализацией здесь и далее понимается приведение слова к именительному падежу единственного числа для существительного, прилагательного, к инфинитиву для глагола и т.д. Константа задается в XML-файле в виде:

<ARG TEXT="*константа*"/>

Например, константами являются имена улиц, людей, числа, представляющие собой номера домов, квартир, понятия, слова-действия и любые другие нормализованные слова, встречающиеся в текстах ЕЯ.

5.2. ***Тип элемента*** - это указатель класса, к которому относится константа. Типы задаются для определенных объектов и служат для указания, что значит в нем тот или иной элемент. Типы - это выделенные константы. Они вводятся в ЛЗ при настройке процессора.

5.3. ***Типизированная константа*** - это константа с указанием ее класса. Задается в виде:

<ARG TEXT="*константа*" TYPE = "*тип элемента*"/>

Например, если объектом является адрес, то указывается, что данное слово - улица, число - номер дома и т.д. Наборы типов задаются при настройке процессора.

043-Kuznetcov_I:_Layout 1 18.05.2008 15:37 Page 288

5.4. ***Атрибут*** - это константа, характеризующая свойство объекта. Задается в виде:
<ARG TYPE = "*атрибут*"/>

5.5. ***Ссылка на объект***. Каждый объект имеет свой уникальный номер, называемый ***идентификатором***. Ссылка на объект задается в виде:
<ARG REF = "*идентификатор объекта*"/>.

5.6. Компонента XML-файла, называемая ***объектом*** (или просто объект), определяется идентификатором, типом и содержит упорядоченное множество элементов, каждое из которых есть или константа, или свойство, или ссылка на другой объект, называемый ***дочерним***. В конце дается описание объекта - текстовый фрагмент, на основе которых был сформирован данный объект.

Тип объекта - это выделенная константа. Такие константы задаются при настройке процессора: для каждого типа объектов - своя константа (*FIO, DATE, ADDRESS* и др.). Один объект может быть дочерним по отношению к нескольким объектам. Два ограничения - отсутствие циклической зависимости и объект не может ссылаться на действия. Наоборот, действия ссылаются на объекты, см. п.5.7. Объект задается в виде:

```
<OBJECT ID="идентификатор" TYPE="тип объекта">
   <ARG ... />
   <ARG ... />
   :
  <SOURCE>описание объекта</SOURCE>
</OBJECT>
```

Здесь <ARG ... /> - или константа, или свойство, или ссылка на другой объект. Порядок элементов в объекте определяется порядком соответствующих слов или фрагментов в тексте, на основе которых был сформирован объект.

5.7. Компонента XML-файла, называемая ***действием*** (или просто действие), определяется идентификатором, типом (соответствует глаголу) и содержит упорядоченное множество элементов действия, каждое из которых есть или константа, или ссылка на объект, или ссылка на другое действие. Подобно объектам, действия также могут содержать произвольный неупорядоченный набор атрибутов (свойств). Действие задается в виде:

```
<ACTION ID="идентификатор« TYPE="тип действия">
  <ARG ... />
  <ARG ... />
  :
</ACTION>
```

В отличие от объектов, у действий нет описания. Порядок элементов в действии определяется порядком соответствующих слов или фрагментов в тексте, на основе которых был сформировано действие.

5.8. Компонента XML-файла, называемая ***отношением*** (или просто отношение), определяется типом (именем отношения) и содержит два элемента, каждый из которых это ссылка на объект, действие или константа. Отношение задается в виде:

```
<RELATION TYPE="тип отношения">
  <ARG REF="идентификатор 1-го объекта или действия"/>
  <ARG REF="идентификатор 2-го объекта или действия"/>
</RELATION>
```

Вместо идентификаторов могут быть константы. Фактически отношение - это важный частный случай двух элементного действия, у которого отсутствуют идентификатор и свойства.

5.9. Компонента XML-файла, называемая ***предложением*** (или просто предложение) состоит из упорядоченного набора констант и ссылок на объекты или действия, которые были сформированы на основе соответствующего предложения ЕЯ. В конце дается текст самого предложения, взятого из исходного текста. Предложение задается в виде:

```
<SENTENCE>
   <ARG ... />
   <ARG ... />
   :
```

043-Kuznetcov_I:_Layout 1 18.05.2008 15:37 Page 289

<SOURCE>*исходное предложение ЕЯ-текста*</SOURCE>
</SENTENCE>

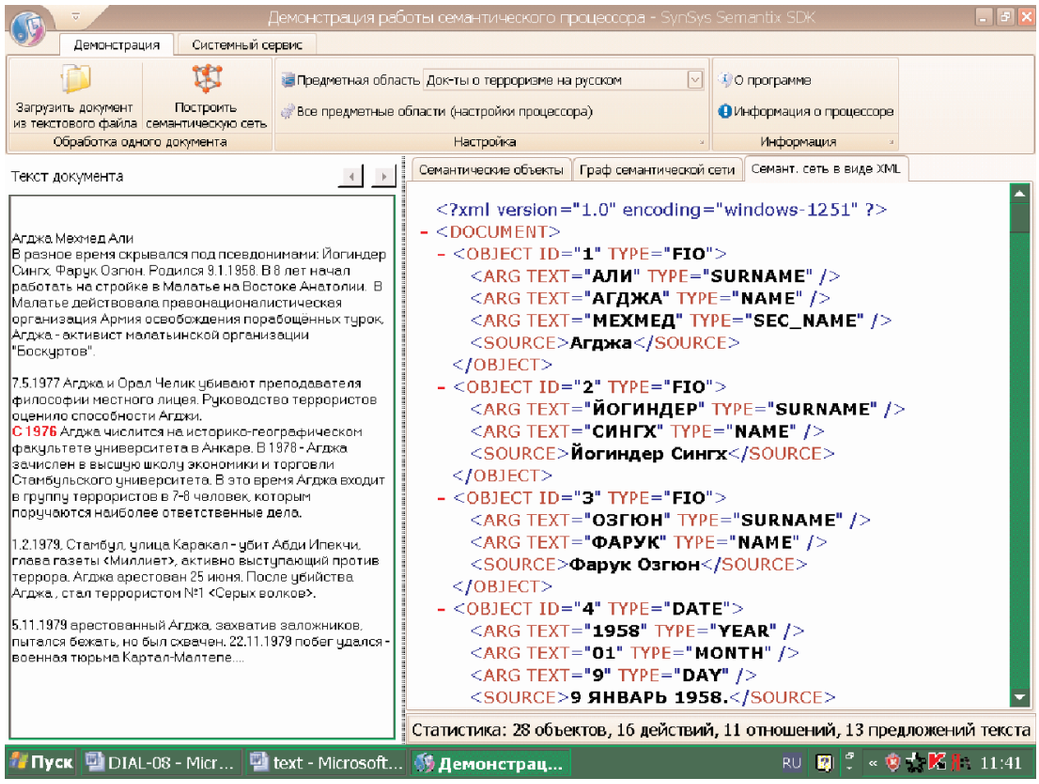
5.10. Выходной ***XML-файл*** состоит из вышеперечисленных компонент и имеет вид:

<?xml version="1.0" encoding="windows-1251" ?>
<DOCUMENT DOC_NUM="*номер документа*">
<OBJECT ...>
содержимое 1-го объекта
</OBJECT>
<OBJECT …>
содержимое 2-го объекта
</OBJECT>
:
<ACTION …>
содержимое 1-го действия
</ACTION>
:
<RELATION …>
содержимое 1-го отношения
</RELATION>
:
<SENTENCE *содержимое 1-го предложения* >

</SENTENCE>
:
</DOCUMENT>

Порядок предложений XML-файле соответствует их порядку в исходном тексте.
Пример XML-файла представлен на рис.4.

Рис.4. Пример XML-файла для представления семантической структуры



043-Kuznetcov_I:_Layout 1 18.05.2008 15:37 Page 290

На Рис.4 имеются типизированные константы для объектов FIO (*лицо*), DATE (*дата*). При этом видно, что описание объекта не всегда берется из текста. Если процессор по элементам объекта в указанном интервале (задается средствами позиционирования предложения) не может найти нужное описание, то процессор формирует свое описание, как например*, 9 январь 1958* вместо *9.1.1958.*.

В XML-файле имеются все компоненты, необходимые для различных приложений. Нормализованные элементы являются основой организации различных видов «объектного» или семантического поиска. Описания служат для построения различного рода досье, отчетов, форм и т.д. XML-файлы могут быть основой для автоматического построения RDF- представлений. Это перспективное направление исследований в плане развития объектно-семантического WEB.

Заключение

В настоящее время предлагается версия семантико-ориентированного лингвистического процессора - ***Semantix 1.0***, обрабатывающего документы в различных предметных областях на русском и английском языках. Качество работы процессора может оценить любой пользователь на своих документах, выйдя на сайт [16].

Semantix 1.0 представляет собой библиотеку COM-объектов и функций, предназначенную для автоматической обработки текстов естественного языка- русского и английского. Модульная структура Semantix позволяет без больших трудозатрат встраивать его в системы обработки текстовой информации, например, системы документооборота, электронные издания и т.п. Представляется также перспективным использование Semantix 1.0 как основу организации баз знаний и разичного вида семантических (объектных) поисков.

Список литературы

1. Кузнецов И.П. Семантические представления // М.: Наука. 1986г. 290 с.
2. Кузнецов И.П., Мацкевич А.Г. Семантико-ориентированные системы на основе баз знаний. Монография. М.Связьиздат. 2007. 173 с.
3. Cunningham, H. Automatic Information Extraction // Encyclopedia of Language and Linguistics, 2cnd ed. Elsevier, 2005.
4. Han J. and Kamber, M. Data Mining: Concepts and Techniques // Morgan Kaufmann, 2006.
5. FASTUS:a Cascaded Finite-State Trasducerfor Extracting Information from Natural-Language Text. // AIC, SRI International. Menlo Park. California, 1996.
6. Кузнецов И.П. Методы обработки сводок с выделением особенностей фигурантов и происшествий // Труды международного семинара Диалог-1999 по компьютерной лингвистике и ее приложениям. Том 2. Таруса 1999.
7. Кузнецов И.П., Мацкевич А.Г. Система извлечения семантической информации из текстов естественного языка // Труды международной конференции Диалог 2001 по компьютерной лингвистике и её приложениям: Т.2. М.: Наука 2002.
8. Кузнецов И.П., Особенности обработки текстов естественного языка на основе технологии баз знаний // Сб. ИПИ РАН, Вып.13, 2003 г. стр. 241-250.
9. Kuznetsov, I., Kozerenko, E. The system for extracting semantic information from natural language texts // Proceeding of International Conference on Machine Learning. MLMTA-03, Las Vegas US, 23-26 June 2003, p. 75-80.
10. Кузнецов И.П., Мацкевич А.Г. Англоязычная версия системы автоматического выявления значимой информации из текстов естественного языка // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2005», Звенигород, 2005.
11. Кузнецов И.П., Мацкевич А.Г. Семантико-ориентированный лингвистический процессор для автоматической формализации автобиографических данных // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2006», Бекасово, 2006, стр. 317-322.
12. Кузнецов И.П., Сомин Н.В. Англо-русская система извлечения знаний из потоков информации в Интернет-среде. // Сб. ИПИ РАН, Вып.17, 2007, стр. 236-253.
13. Кузнецов И.П., Мацкевич А.Г. Лингвистические и алгоритмические аспекты выделения объектов и связей из предметно-ориентированных текстов // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2007«, Бекасово, 2007, стр. 333-342.
14. Сомин Н.В., Соловьева Н.С.., Шарнин М.М. Система морфологического анализа: опыт эксплуатации и модификации // Системы и средства информатики, Вып. 15 // ИПИ РАН - М.: Наука, 2005. - с. 20-30.
15. Ермаков А.Е.. Автоматическое извлечение фактов из текстов досье. Опыт установления анафорических связей // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2007«, Бекасово, 2007, стр. 172-177.
16. ДЕМО-версия процессора Semantix - http://www.semantix4you.com

043-Kuznetcov_I:_Layout 1 18.05.2008 15:37 Page 291

Приложение

Пользовательские настройки (выбор объектов и правил) для области «*Документы о терроризме на русском*».

1. ФИО лиц	2. Арабские ФИО	3. Идентификация лиц
4. Клички	5. Приметы	6. Даты, время
7. Интервалы времени	8. Телефоны	9. Телефоны из зап.книжек
10. E_MAIL	11. WEB-сайт	12. Место. Адрес
13. Организации	14. Работа, должность	15. Оружие
16. Автосредства	17. Террористы	18. Вооруженные силы
19. Номерные вещи	20. Паспорт, документы	21. Национальность
22. Номера счетов, ИНН	23. Наркотики	24. Значимые объекты
25. Товары	26. Службы МВД	27. Статьи УК
28. Крим. Дело	29. Выделение примет	30. Пользовательские объекты
31. Свойства объектов	32. Приметы	33. Идентификация местоимений
34. Словосочетания	35. Числовые показатели	36. Однородные члены
37. Термины	38. Синонимы	39. Действия, события.

044-Kuznetsova_AI:_Layout 1 18.05.2008 15:39 Page 292

СИНКРЕТИЗМ ЧАСТЕЙ РЕЧИ В РУССКОМ И УРАЛЬСКИХ ЯЗЫКАХ (К ВОПРОСУ О СПЕЦИФИКЕ СОСТАВЛЕНИЯ ДВУЯЗЫЧНЫХ СЛОВАРЕЙ ДЛЯ ЯЗЫКОВ РАЗНЫХ ТИПОВ)

PARTS OF SPEECH SYNCRETISM IN RUSSIAN AND URALIC (THE PROBLEM OF THE COMPOSITION OF BILINGUAL DICTIONARIES FOR INFLECTIONAL AND AGGLUTINATIVE LANGUAGES)

Кузнецова А.И. *(aikuznec@yandex.ru)*
Московский государственный университет им. М.В. Ломоносова

В языках флективных (как русский) и агглютинативных (как уральские языки) существуют разные способы подачи синкретичных частей речи в толковых словарях. Специфика и трудности составления двуязычных словарей тесно связаны со способностью синкретичных слов функционировать в предложении в качестве различных его членов.

Термин *синкретизм*, вынесенный в заглавие, напоминает хамелеона (< греч.χᾰ̃μαι λέων ‘на земле + лев, львиная шкура’): подобно представителю отряда небольших древесных ящериц (из семейства пресмыкающихся), быстро меняющих окраску своей кожи, та или иная часть речи может выполнять функцию другой (а иногда сразу нескольких) частей речи. Переход слова в другую часть речи сопровождается, с одной стороны, изменениями в парадигме слова. Напр., в современном языке существительные утратили древние формы звательного падежа (Г*осподи*! *Боже* мой!) при переходе единичных форм в разряд междометий. Точно так же у глаголов исчезли формы аориста и некоторых других прошедших времен, а сохранившиеся формы превратились в частицы (*бы*)[1] и т.д. С другой стороны, эти изменения обычно бывают связаны со значительными изменениями в самом функционировании слова, при котором слово начинает выполнять нередко новые синтаксические функции. Иными словами, одна и та же синтаксическая функция осуществляется разными частями речи[2]. Так, существительное наряду с выполняемой им функцией подлежащего или дополнения начинает функционировать как сказуемое (предикат) или как обстоятельство и т.п. Едва ли не самым активным в этом процессе является в русском языке местоимение – оно встречается в функции определения, дополнения, подлежащего, предиката, обстоятельства; глагол также не ограничивается своей основной функцией сказуемого (предиката), а может быть и подлежащим и т.д. Это явление перехода одной части речи в другую и выполнение ею разных функций было давно отмечено исследователями русского языка, причем все попытки теоретического осмысления этого факта породили великое множество терминов. Наряду с термином *синкретизм* для передачи тех явлений, о которых пойдет речь, употребляются термины *контаминация, промежуточные факты, конверсия, кумуляция, частеречная омонимия, гибридность* и др. В 1960-1970-е гг. много говорилось также о центральных и периферийных единицах в языке, с которыми связаны перечисленные выше термины. Акад. В.В. Виноградов, пользуясь термином *гибридность, гибридный, гибридизация*, писал о совмещении в одной единице свойств различных структурно-семантических классов, о «грамматическом перерождении» одной части речи в другую [2, с.418]. Встречаются и другие термины для анализируемого явления, но, как справедливо замечал А. Мартине, «важно не ставить на явлениях определенные этикетки, а наблюдать и объяснять процесс» [3, с.35].

Именно такую попытку в свое время предпринял А.Я.Баудер [4, с.85ff.], который показал (следуя работам В.В. Бабайцевой [5], [6], В.В. Виноградова и др.) суть происходящих при гибридизации процессов на конкретных примерах существительных и прилагательных, продемонстрировав процесс преобразования

[1] Нередко после «перехода» какой-либо формы (или нескольких форм) слова в другую часть речи оставшаяся парадигма превращается в неполную, а ее члены идентифицируются как фразеологические единицы, напр., *вверх тормашки и вверх тормашками* (при исчезновении форм *тормашка / тормашки*). Слово *тормашка* у Даля объясняется как отглагольное существительное (от глагола *тормошить*:. *(вверх) тормашки полетел* означает ‘перекувыром, через голову, вверх ногами. Полетел от Машки *вверх-тормашки*!’ [1]). Этимология слова неясна (Ср. польск. tarmosić). Ср. также: *на закорки / на закорках, на карачках / на карачки* и др. при наличии в говорах формы *карач* ‘колени’ и т.д.

[2] В данной работе не рассматриваются омоморфы, при появлении которых происходит совпадение слов разных частей речи обычно в какой-либо одной единственной форме (редко в двух); напр., слова *пил, пила* или *дул, дула* – это глаголы прош. времени и сущ. в Gen.Pl и Nom.Sg или Gen.Sg; *сев* (Nom.Sg сущ. и деепр.).

044-Kuznetsova_AI:_Layout 1 18.05.2008 15:39 Page 293

прилагательных в существительные как двусторонний процесс. Это процесс взаимодействия двух классов: с одной стороны, прилагательное утрачивает исходный признак прилагательного; с другой, слово, не утратив своего признака, приобретает свойства класса существительного. В какой-то момент количество сосуществующих признаков может колебаться, появляются омонимы («периферийные и ядерные субстантиваты», по терминологии автора). Периферийные субстантиваты способны сочетаться с существительными и одновременно могут функционировать без существительных (напр.: *штатский человек* и *штатский, снотворное лекарство* и *снотворное* и др.). Ядерные же субстантиваты всегда употребляются без существительных: *лесничий, вожатый, набережная, существительное, прилагательное* и др. (Последнее слово не случайно обыгрывается у Д.И.Фонвизина). В результате по характеру совмещенных признаков А.Я. Баудер выделяет ступени (которые могут быть оспорены), демонстрирующие постепенный переход слов от прилагательных к существительным.

Ядерные прилагательные (по А.Я. Баудеру) проходят три стадии, прежде чем достигнут стадии ядерных существительных (при этом некоторые до третьей стадии так и не доходят). Данный переход прилагательного в разряд существительных можно схематически изобразить следующим образом:

А (ядерные прилагательные) → **Аб** (окказиональные субстантиваты, которые семантически не закреплены и структурно не оформлены: *главный врач* (*или бухгалтер, инженер, научный сотрудник*) → **АБ** (синкретичные субстантиваты, которые семантически закреплены, но структурно не оформлены: *любимый, милая, усатый, лысый, косой, хромой, рабочий*) → **аБ** (периферийные субстантиваты закреплены семантически и структурно оформлены: *штатский* или *военный* (*человек*); *сердечное, слабительное, противогриппозное* (*лекарство*)*; докладная, объяснительная* (*бумага*); *скорый, почтовый, пассажирский* (*поезд*) → **Б** ядерные субстантиваты (*Киев, Соловьев; тепло, добро; портной; вселенная; мостовая; существительное; пирожное, жаркое; животное, сохатый* [4, с.95]. Промежуточные стадии могут быть иными, но суть самого процесса остается: прилагательные в современном русском языке переходят в другую часть речи и начинают выполнять новые функции. В ССРЛЯ [7] примеры типа **АБ** и **аБ** даются с пометой «в знач. сущ.».

Переходы одной части речи в другую наблюдаются в русском языке с древних времен и известны не только между существительными и прилагательными, но и среди многих знаменательных и служебных частей речи. Так, древнерусские глагольные формы переходили в частицы (*бы, де*), в междометия (*чу*!); именные – в междометия (*Господи! Исполать*![3] */ исполати*!), в наречия (*до́ма, внизу*) и т.п. В ходе истории эти процессы иногда ограничиваются одним переходом, а порою могут быть продолжены: так, глагольная форма переходит в междометие, а позднее междометие может превратиться в какую-то иную часть речи, напр., во вводное слово (именно так возникло просторечное слово *чу*). В древнерусском языке формы местоименных (полных) и кратких прилагательных могли постепенно переходить в разряд существительных; последние в современном языке этот статус в одних случаях утратили при сохранении статуса прилагательных, в других сохранили его. В ССРЛЯ приводятся современные слова *погребной* ‘относящийся к погребу; свойственный погребу (напр., погребная сырость)’ и *подвальный* ‘относящийся к подвалу; свойственный подвалу’. Первое из этих слов в древности имело также значение ‘тюремный’ (отмеченное в СРЯ XI-XVII в. [8] и у И.И. Срезневского, правда, в форме *погребьныи*) и значение ‘тот, кто ведает погребами’, выступая в этом случае в функции существительного. Второе слово (*подвальный*) в приведенном выше значении в СРЯ XI-XVII вв. (при наличии слова *подвалъ*) отсутствует вообще[4]. В современном русском языке *погребной,* сохранив свое первичное значение, утратил значение ‘тюремный’. Слово же *подвальный* помимо указанного первичного значения приобрело в современном языке значение ‘рабочий, занимающийся упаковкой и хранением вин’ (перейдя при этом в класс существительных) и значение ‘размещенный в подвале газеты: *подвальная* статья (спец.)’. В СРЯ XI-XVII вв. дается большое количество синкретичных слов, подобных приведенным выше и с пометой, в значении какой части речи слово употреблено. Среди них преобладают слова, имеющие форму и значение прилагательных и вместе с тем означающие разного рода налоги, подати, пошлины, поборы, платы, будучи используемы в качестве существительных: *ловчее* (род налога[5])*, поголовный,-ая,-ое; подворный, -ое* и *подворенный, -ое, подлазное*, *подушный*, *полозовый, угловой.* Среди слов, функционирующих двояко (как существительное и прилагательное), немало слов имеет значение ‘человек, живущий, находящийся где-то’, напр., *подводский*, *подземный, подморский.* Нередки также прилагательные, означающие человека по роду занятий (*ловчий* ‘охотник, рыболов’ и ‘приученный к охоте на кого-л.’. К тем же двум частям речи относятся слова *охочий*, *пѣвчий; писчий,*

[3] Данное слово, как правило, уже не известное молодым, в ССРЛЯ зафиксировано с пометой устар. в значении ‘Хвала, слава (в обращении как выражение одобрения, восхищения)’. Там же приводятся примеры из И.А.Крылова и А.К.Толстого.

[4] Слово *подвальный,* данное в СРЯ XI-XVII вв., образовано от существительного *подвалье,* означающего ‘подводная пещера и глубокое озеро карстового образования (как рыболовное угодье)’; пример XVII в. [8].

[5] В СРЯ XI-XVII вв. примеры на употребление прилагательного *ловчее* (ср.р.) в качестве существительного приводятся, начиная с XIII в.

044-Kuznetsova_AI:_Layout 1 18.05.2008 15:39 Page 294

подвойский, подвластный и т.п.). Встречаются примеры и на совпадение в одной форме не только прилагательных и существительных, но и существительных и наречий, наречий и предлогов и т.д. Синкретичными могут быть междометия и наречия, как слово *исполать* (из греч. Εις πολλα έτη 'на многие лета' (междометие), а во фразе *исполати сии образъ написанъ* слово *исполати* – наречие со значением 'навечно').

Следует обратить внимание на то, что большинство из только что приведенных и давно исчезнувших из современного литературного и разговорного языка слов широко употребляются в современных лингвистических и исторических работах. (См., напр., историческое исследование о жизни Разрядного приказа в 1676/77 г. О.В. Новохатко [9, с.622-637]). Примеры, приводимые в работах такого рода, свидетельствуют о том, что синхрония «в чистом виде» – иллюзия: каждый синхронный срез имеет в своей структуре элементы и прошедших, и будущих систем. Как писал в свое время Ю.Н. Тынянов, «система не есть равноправное взаимодействие всех элементов, а предполагает выдвинутость группы элементов ('доминанта') и деформацию остальных» [10, с.277]. Данное высказывание касается не только литературных произведений (которые имел в виду Ю.Н. Тынянов), но и разных языковых страт (пластов) в современном русском языке.

До сих пор речь шла о синкретичных частях речи (по преимуществу, об именах) в русском языке. При этом частеречный синкретизм наблюдается не только в пределах знаменательных частей речи, но и между знаменательными и служебными словами. Возможность употребления слова в качестве другой части речи означает, что обычно слово начинает функционировать как другой член предложения. Части речи, по терминологии Э. Сэпира, являются основными понятиями в отличие от членов предложения, которые он переименовал в «чисто реляционные понятия». Не исключено, что зависимость может быть обратной: слово употребляется как другой член предложения и благодаря этому у него появляется омоним – другая часть речи. Судя по толковым словарям (напр., [7]), датирующим приводимые примеры, второй вариант имеет больше сторонников, т.е. функция того или иного слова важнее, чем факт отнесения его к той или иной части речи. Развитие членов предложения (особенно второстепенных) переплетается с изменениями в функционировании частей речи. Оба процесса тесно взаимосвязаны. По словам Е.И. Янович, «трудно говорить о том, как формировались средства выражения каждого второстепенного члена предложения в отдельности. Этим и вызывается необходимость рассмотреть вначале исторические изменения в области приименных второстепенных членов, выяснив основные тенденции в развитии определения и отчасти дополнения, а затем рассмотреть средства выражения приглагольных второстепенных членов, заметив основные линии развития дополнения и обстоятельства» [11, с.254]. Точно так же есть ограничения, связанные с глаголом, который может выполнять функцию далеко не всех членов предложения. Тем не менее, отдельные глагольные формы (напр., у глагола *знать*) уже давно перешли в разряд служебных слов и в качестве таковых встречаются в функции вводного слова (*знать* 'видимо, вероятно': «*знать, забило сердечко тревогу…*»). Формы от глагола *пустить* превратились в союз и во вводное слово с несколькими значениями (*пусть*); форма повелительного наклонения глагола *быть* может функционировать как частица («*Да будь я и негром преклонных годов…*»). Наречие *просто* может употребляться также как частица (усилительная и ограничительная) и т.д.

Появление синкретичных частеречных форм в русском языке – процесс длительный, начавшийся в древнерусском языке и продолжающийся и в настоящее время. Анализ описанного явления в современном языке позволяет проследить постепенность перехода слов из одной части речи в другую, установить наличие в языке переходных форм, что значительно труднее делать по текстам прошлых веков. Характерной чертой частеречного синкретизма в русском языке является, с одной стороны, лишь частичное пересечение слов разных частей речи в одной или нескольких формах и расхождение в большинстве других. Исключение представляют синкретичные формы существительных и прилагательных, падежные формы которых совпадают (иногда в пределах одного числа), но синтаксические функции их различны. Однако синкретизм существительного с другими частями речи (напр., с наречием) обычно наблюдается в одной или нескольких формах, причем часто в аналитических формах с предлогами / префиксами (напр., Род.п. существительного *дома* (*возле дома*) и наречие *дома* (*его нет дома*); *на встречу в школу пришли все выпускники,* но *иду навстречу ему*; *в самый низ* (чего-то) и *он смотрит вниз*).

Явление частеречного[6] синкретизма, встречающееся в русском языке, широко распространено и в ряде финно-угорских языков, но выглядит оно обычно иначе, чем в русском языке, что связано с разницей грамматического строя (русский – флективный язык, а уральские языки – агглютинативные). В частности, это зависит от разницы в типах предложений: SVO–обычная структура русского предложения, а SOV – уральских языков.

[6] Существует и другой вид синкретизма, названный в 1929 г. С.И. Карцевским *асимметричный дуализм* структуры знаков, при котором разные означаемые (грамматические значения), выполняющие несколько функций, представлены в одной форме [12, с.85-90], как, напр., латинский Abl. передающий функции Instr. и Loc. Разновидности синкретизма такого типа в самых разных его проявлениях посвящена монография М. Баермана, Л. Брауна и Г.Г. Корбета 2005 г. издания, подробно отрецензированная П.М. Аркадьевым в журнале «Вопросы языкознания» [13, с.123-129].

044-Kuznetsova_AI:_Layout 1 18.05.2008 15:39 Page 295

Частеречный синкретизм в уральских языках проявляется двояко. В самодийских языках и в мордовских имя может спрягаться в 1-2 лицах Sg-Pl (в самодийских языках и в Du) в настоящем времени, в чем, собственно говоря, и проявляется функционирование существительного в качестве предиката. В 3л. имя не получает глагольного окончания (в отличие от 1-2 лиц). Это явление, называемое в указанных языках вербальной репрезентацией (наряду с данным существуют и другие термины), наблюдается, напр., в селькупском (тазовско-туруханский северный диалект) и в эрзянском языке (в частности, в идиоме села Шокша).

Селькупский язык Mat qum-**ɔ̄k** ‘я человек’ (букв. ‘я человек-аю’, т.е. ‘я человек есмь’)

Tat qum-**ɔ̄nty** ‘ты человек’ (букв. ‘ты человека-ешь’)

Но: T&p qum-∅ ‘он человек(ает)’ (и т.д. в прошедшем времени Sg, Du, Pl.) [14,с. 188-190].

При адъективной (элевой) репрезентации существительных их формы входят в парадигму существительного, но синтаксически используются как определение. Так, сочетание qorqyt tary**l’** porqy не ‘медведя шерстяная одежда’ (букв.), а ‘одежда из шерсти медведя’; pó**l’** tol’čy**l’** mytyn ‘мазь для деревянных лыж’ (букв. ‘деревянная лыжная мазь’). При некоторых глаголах существительное в адъективной форме выполняет функцию дополнения: Qomtä**l’** pérysa ‘ (Он) деньги (букв. ‘денежное’) искал’; qorsa**l’** účak ‘я теслом (букв. ‘тесловым’) работаю’ и т.д. Иными словами, одна и та же форма на **-l’,** свойственная и существительному, и прилагательному, функционирует как определение и дополнение.

Эрзянский язык (с.Шокша):

Mon Sokaws’an er’t’-**an**, ton Moskow ošin’-**at**, a son Kraut’a-∅

‘я житель(≈жительствую) Сокаевский, ты – города (≈го́родишься) Москвы, он – Кураева’.

Min’ miRh’t’ava-**tama**: mon mird’-**an**, ton kozejk-**at**

‘Мы супруги (букв.: Мы муж-жена < mir’d’e + ava): я муж, ты – жена’ (букв.’хозяйка’).

В других самодийских и в мокшанском языках также существует частеречный синкретизм подобного типа. Интересно при этом отметить, что в рассматриваемом идиоме с. Шокша, как и в обоих мордовских языках, предикативные формы возможны не только от имен, но и от наречий, местоимений, числительных, что отмечали лингвисты еще в середине прошлого века. Напр.: **эрз**. *Петя тосо, мон тес**ан*** «Петя там, я здесь (нахожусь)»; *мон тесэлинь* «я здесь был» [15, с.114-115]; *од**ат*** «ты молодой (еси)»; *котоц**ят*** «ты шестой»; *васол**ят*** «ты далеко (еси)», *ягань вакс**сат*** «ты около моего друга (находишься)»; **мокш.** *Тя мон**ан.** Панчк**!*** «Это я. Открой!» и др. [16, с.169].

Предикативные формы, о которых идет речь, образуются (за редким исключением) от одушевленных существительных (наименования профессий, жителей, родственников, национальности субъектов и т.п.), составляющих своеобразный криптокласс в названных языках. (Подробнее о связи лексико-семантических классов слов с грамматическими категориями см. в работах Б.Л.Уорфа [17], О.О. Борискиной & А.А. Кретова [18], А.И. Кузнецовой [19], Е.А. Цыпанова [20] и др.). При наличии сходства между русским и названными уральскими языками в самом существовании частеречного синкретизма, проявляющегося в использовании одной и той же части речи (грамматической формы имени) в функции сказуемого и подлежащего, можно указать и на их различие.

В русском языке, где имеется синкретизм существительного и прилагательного, эти части речи обычно в функции предиката выступают в именительном или творительном падежах (об их семантическом противопоставлении см. [Гиро-Вебер 21]) независимо от того, идет ли речь о 1-2-ом или 3-ем лицах существительного: *я / ты / он студент.* В нескольких уральских языках (см. выше) при частеречном синкретизме существительное и прилагательное функционируют как сказуемое без показателя только в 3 л., а в 1-2 оформляются глагольными показателями соответствующих чисел. Однако частеречный синкретизм присутствует и в других уральских языках, но форма проявления у него иная, точно так же, как втянутым в этот процесс оказывается иной круг частей речи.

В этом втором своем проявлении частеречный синкретизм встречается в большем количестве языков. В них наблюдается употребление одного и того же слова в разных синтаксических функциях без использования парадигм других частей речи, как это было в самодийских и мордовских языках. Так, в языке бесермян, коми и в марийском изменение синтаксической функции слова происходит только при изменении порядка слов. В результате изменения порядка слов в сочетании одно и то же слово может выступать то как подлежащее, то как сказуемое, определение или обстоятельство. Напр.:

бес. *amsor* d’er’em ‘*узкая* рубаха’ и d’er’em *amsor* ‘рубаха *узка*’; m‘i *načar* uleškom ‘мы *бедно* живем’ и *načar* korka ‘*бедный* дом’ [22]; **эрз.** Кудосонок пальсь *валдо* ‘В доме-нашем горел *свет*’ и *Валдо* весь прядовсь *‘Светлая* ночь кончилась’**[7].**

[7] В **к.-з.** часты омоморфы сущ. или прил. в исходных формах и формы повелительного наклонения глагола: *потш* ‘жердь’ и ‘загораживай!’ (<потшны); *сыв* ‘талый; жир’ и ‘расплавляй!’ (< сывны); *тіль* вӧр ‘густой лес, чаща’ и синтӧ эн *тіль* ‘глаза не *три*!’ (ср. аналогичные случаи в русском языке; см. примечание 2).

044-Kuznetsova_AI:_Layout 1 18.05.2008 15:39 Page 296

Описанное своеобразие русского и некоторых уральских языков в области частеречного синкретизма заставляет задуматься о том, как разумнее подавать материал не только в толковых одноязычных, но и в двуязычных словарях. В толковых словарях уральских языков, входящих в состав России, принято (хотя последовательно это не соблюдается) при толковании на русском языке давать все синкретичные формы в одной словарной статье. В СРЯ XI-XVII вв. следуют тем же правилам подачи материала. Вместе с тем в ССРЛЯ наречия, глаголы даются от имен изолированно, а подача существительных и прилагательных не унифицирована: в одних случаях в пределах этих статей ставится помета «в знач. прил.» (или сущ. соответственно), в других синкретичное слово выносится в самостоятельную словарную статью. Такая же непоследовательность наблюдается и с частицами, союзами, вводными словами, междометиями. Очевидно, что, если учитывать наличие гибридных форм у существительных, прилагательных и других частей речи, то такая подача могла бы быть оправдана, но для этого надо проделать огромную предварительную работу, обнаруживая гибридные формы всех частей речи. Когда же тот же вопрос касается двуязычных (переводных) словарей, то сложности усугубляются, и вопрос остается открытым.

Список литературы

1. Даль В.И. Толковый словарь живого великорусского словаря. – М., 1995.

2. Виноградов В.В. Русский язык. Грамматическое учение о слове. – М., 1947. 784с.

3. Мартине А. Принцип экономии в фонетических изменениях (проблемы диахронической фонологии). – М., 1960. 262 с.

4. Баудер А.Я. Части речи – структурно-семантические классы слов в современном русском языке. – Таллин, 1982. 184с.

5. Бабайцева В.В. Гибридные слова в системе частей речи современного русского языка // Русский язык в школе. – М, 1971, №3, с.81-84.

6. Бабайцева В.В. Зона синкретизма в системе частей речи современного русского языка. // Филологические науки. – 1983, №5, с.35-42.

7. ССРЛЯ – Словарь современного русского литературного языка. Т.1-17. – М., 1947-1965.

8. СРЯ XI-XVII в. – Словарь русского языка XI-XVII веков. – М., 1978-

9. Новохатко О.В. Разряд в 185 году. – М., 2007. 640 с.

10.Тынянов Ю.Н. О литературной эволюции // Поэтика. История литературы. Кино. – М., 1977 (1927), с.270-281.

11.Янович Е.И. Историческая грамматика русского языка. – Минск, 1986. 320с.

12. Карцевский С.И. Об асимметричном дуализме лингвистического знака // Звегинцев В.А. История языкознания XIX-XX вв. в очерках и извлечениях. Ч.II. – М., 1965, с.85-90.

13. Аркадьев П.М. (рец. на книгу: M.Baerman,D. Brown, G.G. Corbett. The syntax-morphology interface. A study of syncretism. Cambridge, 2005) // ВЯ, 2006, №5, с.123-129.

14.Очерки – Кузнецова А.И., Хелимский Е.А., Грушкина Е.В. Очерки по селькупскому языку. Тазовский диалект. Т I. – М., 1980.

15.Колядёнков М.Н. Грамматика мордовских (эрзянского и мокшанского) языков. Ч.II. Синтаксис. – Саранск, 1954. 327с.

16.Цыпайкина В. Об особых средствах выражения темпоральных отношений в мордовских языках // FU X. Pars II. Linguistica. – Joshkar-Ola, 2005, p.169.

17.Уорф Б.Л. Грамматические категории // Принципы типологического анализа языков различного строя. – М., 1972, с.44-60.

18.Борискина О.О., Кретов А.А. Теория языковой категоризации. Национальное языковое сознание сквозь призму криптоклассов. – Воронеж, 2003. 211с.

19.Кузнецова А.И. Категория вербальной репрезентации в уральских языках // Конференция по уральским языкам, посвященная 100-летию К.Е. Майтинской. – М., 2007, с.120-126.

20. Цыпанов Е.А. Особенности обозначений понятий молодой – старый в финно-угорских языках // Конференция по уральским языкам, посвященная 100-летию К.Е. Майтинской. – М., 2007, с.247-252.

21.Гиро-Вебер М. Существительное в функции именного сказуемого в современном русском языке: возможно ли еще говорить о семантическом противопоставлении «Им. vs. Тв.»? // ВЯ, №1, 2007, с.18-26.

22. ПМА – Полевые материалы автора

045-Kustova_G:_Layout 1 18.05.2008 15:43 Page 297

О «НЕНОМИНАТИВНЫХ» ЭЛЕКТРОННЫХ СЛОВАРЯХ*
ABOUT «NON-NOMINATIVE» DICTIONARIES (LEXICAL DATABASES)

Кустова Г.И. *(galina03@mtu-net.ru)*
Московский государственный педагогический университет

В статье предлагается проект электронного словаря, включающего обстоятельственные обороты и грамматические единицы, возникшие на базе падежных и предложно-падежных форм существительных (*на ходу; в ходе; полным ходом*) и не имеющие номинативной начальной формы.

В синтаксисе традиционно противопоставляются управляемые формы (реализующие валентные связи) и свободно присоединяемые (разного рода модификаторы). Свободно присоединяемыми распространителями бывают существительные в косвенных падежах с предлогами и без предлогов (так называемое падежное примыкание), наречия, прилагательные, инфинитивы. В данной работе мы будем касаться, в основном, предложно-падежных оборотов. Второе ограничение материала связано с синтаксической функцией. Среди свободно присоединяемых форм встречаются определения (*юбка в клетку*; *пирожок с яблоками*; *родственники из Тулы*) и обстоятельства. В статье речь пойдет только об обстоятельствах.

Управляемые формы выбираются из парадигмы в соответствии с требованиями управляющего слова. При этом они сохраняют семантическое тождество с другими членами парадигмы, т.е. имеют одно лексическое значение, а падежные различия не являются семантически значимыми (синтаксические падежи по Е. Куриловичу). В словаре вся парадигма представлена начальной формой – именительным падежом. Такой способ словарного представления можно условно назвать «номинативным». Свободно присоединяемые формы с обстоятельственной семантикой выступают в виде «адвербиалов» – наречных падежей (в смысле Куриловича) и предложно-падежных групп (в силу большей или меньшей адвербиализованности эти формы уже нельзя считать полноценными падежами, т.е. формами существительного).

Для краткости такие формы будем называть обстоятельственными оборотами. Обстоятельственные обороты обособляются от именной парадигмы и функционируют как более или менее самостоятельные единицы с особым «адвербиальным» значением, которое неотделимо от формы и конструкции. Разумеется, адвербиальные падежные формы сохраняют определенную семантическую связь с существительным, но это такая же связь, как между членами словообразовательного гнезда, куда входят родственные слова разных частей речи.

Можно выделить несколько типов единиц («неноминативных» моделей), которые возникли в результате «отрыва» падежных и предложно-падежных форм от именной парадигмы.

(а) Падежные и предложно-падежные формы, которые функционируют как эквиваленты наречий: *летом*, *боком*, *на ходу*, *с иголочки*, *на скаку*, *с размаху*.

(б) Именные производные предлоги, которые образуются за счет грамматикализации сочетания простых предлогов с абстрактными существительными *вид, масштаб, результат, сила, случай, ход* и др.: *в виде* (*исключения*), *в силу* (*привычки*), *по ходу* (*дела*), *в ходе* (*следствия*), *на случай* (*дождя*), *за счет* (*резервов*) и под.

(в) Обстоятельственные обороты с прилагательными. Эта группа обычно выпадает из поля зрения исследователей, поэтому на ней мы остановимся более подробно.

Обстоятельственные обороты с прилагательными образуются, в основном, на базе тех же существительных, от которых образуются производные предлоги. Если именной предлог управляет падежом (*в виде, в границах, в масштабах, по мере, в отношении, в порядке, в результате, в силу, в случае, за счет, в ходе* ЧЕГО; *в качестве, в роли* КОГО/ЧЕГО), то адъективный обстоятельственный оборот представляет собой своего рода рамку (*в … виде*; *в … порядке* и т.п.), в которую «вставляется» прилагательное или местоимение-прилагательное (*в готовом виде*; *в прежнем качестве*; *во всяком случае*). В результате образуются своего рода пары: *в электронном виде – в виде файла*; *в масштабе страны – в государственном масштабе*; *в порядке эксперимента – в рабочем порядке*; *в случае удачи – в лучшем случае* и т.п. (более подробно такие пары мы рассматриваем в другой работе, см. [2]). Не все адъективные обороты имеют «пару» в виде предлога – некоторые употребляются только в «адъективном варианте» (*в высшей степени*; *в полной мере*). И не все

* Работа выполнена при поддержке РГНФ, проект № 08-04-00183а.

045-Kustova_G:_Layout 1 18.05.2008 15:43 Page 298

обстоятельственные обороты включают предлог – есть беспредложные, образованные на базе творительного падежа, ср. *полным ходом*, *первым делом*, *таким образом*.

Общим свойством падежных и предложно-падежных адвербиалов является «**неноминативность**» как результат утраты связи с именной парадигмой.

(а) Например, наречный оборот *на ходу* имеет значение «не останавливаясь, не прекращая ходьбы» (*ест на ходу*), «во время движения, не дожидаясь остановки» (*спрыгнул на ходу*). Это значение реализуется именно в данном обороте (ср.: **ел* / *спрыгнул в ходе*; **ел* / *спрыгнул во время хода*), поэтому, вообще говоря, не может быть представлено в словаре словом *ход* (в МАС'е это значение представлено как одно из лексических значений слова *ход* – 'перемещение человека или животного на собственных ногах', – однако никаких других примеров, кроме как с оборотом *на ходу*, там не приводится, чего и следовало ожидать).

Связь с именной парадигмой и субстантивные свойства могут утрачиваться не полностью. Это проявляется, например, в том, что некоторые наречные обороты допускают включение адъективов («адъективные вставки»): *под вопросом – под большим вопросом*; *под носом – под самым носом*, ср. также *на полном ходу*; *на всем скаку*; *со всего размаху*; *в самом начале* (ср., однако: *под мухой – *под большой мухой*, (*шел*) *под руку* (*с дамой*) – **под правую руку* (*с дамой*); *судачат за глаза – *за все глаза*; (*произошло*) *на глазах – *на самых глазах*; (*пошло*) *на пользу – *на большую пользу*). Впрочем, лексический состав таких адъективных вставок крайне беден – в основном, это прилагательные и местоимения со значением высокой степени, что согласуется с наречным смыслом оборотов. Определения, ориентированные на субстантивную семантику существительного, не допускаются (ср. **на стремительном скаку*; **под большим носом*), что подтверждает «неноминативность», адвербиальность подобных единиц.

(б) Та же неноминативность свойственна именным предлогам: *в силу задержки – *сила задержки*, (*лекарство продается*) *в виде порошка – *вид порошка*; (*работал*) *в качестве начальника – *качество начальника*.

(в) Если для оборотов типа *на ходу*, *с ходу*, *на лету*, *в силу*, *в виде*, *за счет*, *в качестве* и под. неноминативность естественна и объяснима, т.к. эти предложные группы полностью обособились от именной парадигмы, полностью (или почти полностью) превратились в наречия или предлоги, то для оборотов с прилагательными дело обстоит иначе. Хотя они выполняют обстоятельственную функцию, внешне они похожи на «нормальные» словосочетания (ср. *в добровольном порядке*). Между тем они тоже имеют неноминативный, «наречный» характер.

Среди таких оборотов встречаются два типа единиц: настоящие фразеологизмы (идиомы) вроде *первым делом*; *полным ходом*; *в первом приближении* и т.п. и своего рода коллокации, когда полузнаменательная обстоятельственная рамка сочетается со знаменательным прилагательным (*в горячем виде* → *горячим*; *в срочном порядке* → *срочно*; *в грубой форме* → *грубо*).

Обстоятельственные рамки не «переводятся» в именительный падеж: получается либо неправильное сочетание (*в горячем виде – *горячий вид*, *в лучшем случае – *лучший случай*), либо сочетание с другим значением (*в порядке проверки* ≠ *порядок проверки*).

Второй их важной особенностью является то, что обстоятельственные рамки *в ... виде*, *в ... случае* и т.д. являются сильно грамматикализованными и лексически вырожденными (фразеологизованными) элементами и не употребляются без прилагательных (мы подробно рассматриваем это свойство в другой работе, см. [3]). Этим они отличаются от адвербиалов типа *на скаку*, *на лету*, которые либо вообще не допускают адъективных вставок (**на всем лету*), либо допускают (*на всем скаку*), но не требуют их.

Теми же свойствами – неноминативностью и обязательностью прилагательного – обладают и идиомы, ср.: *Первым делом нужно кричать «Держи вора» – *Кричать «держи вора» – первое дело*. В таких предложениях речь не идет о «первом деле» в ряду других дел (второго, третьего), *первым делом* – эквивалент наречия *сначала*. Разумеется, свободное сочетание *первое дело* тоже существует (*Первое дело труднее*, *чем второе*; *Первым делом он занимался больше*, *чем вторым*); существует и фразеологизм *первое дело*, – однако семантика сочетания *первым делом* не сводится ни к свободному, ни к фразеологическому сочетанию *первое дело*.

У идиом неноминативность и обязательность прилагательного можно объяснить тем, что процесс фразеологизации полностью завершился. У коллокаций обязательность прилагательного имеет другие причины. Тем не менее, несмотря на некоторые различия, идиомы типа *первым делом* и коллокации типа *в массовом порядке* имеют много общего и, несомненно, должны изучаться вместе, в едином комплексе. Те и другие явно имеют грамматический характер и отличаются от фразеологизмов с уникальными лексическими компонентами типа *к шапочному разбору* или *до морковкина заговенья*.

Рассматриваемые обороты выражают типовые обстоятельственные смыслы, и вовлечение в процессы фразеологизации и грамматикализации именно этих оборотов не случайно. С одной стороны, не случайно использование именно этих абстрактных существительных, многие из которых участвуют также и в

045-Kustova_G:_Layout 1 18.05.2008 15:43 Page 299

формировании именных предлогов. С другой стороны, использование определенных адъективов в составе обстоятельственных оборотов также имеет вполне системный характер. В первую очередь, это местоимения: *такой* (*в таком разе, в таком разрезе, в таком случае, таким образом*), *свой* (*своим ходом, своими силами, своим чередом, в свое время, в свою очередь*); *любой* (*в любом случае*, *любой ценой*; *в любой момент*, *под любым предлогом*) и др.; а также прилагательные: *первый* (*в первую голову, в первую очередь, первым делом, первый раз*), *последний* (*в последнюю очередь*, *до последнего вздоха*, *до последней черты*, *до последней капли*); *лучший* (*в лучшем случае*, *в лучшем виде*); *равный* (*в равной мере*, *равным образом*); *полный* (*в полном смысле слова*, *в полной мере*, *в полном рассудке*, *в полный рост*, *на полном ходу*, *на полном серьезе*, *на полную катушку, в полную силу*) и др.

Таким образом, в обстоятельственных оборотах с прилагательными типа *полным ходом*, *в полной мере*, *первым делом*, *на самом деле*, *изо всех сил*, как и в группах (а) и (б), возникает особое «адвербиальное» значение, которое связано с данной формой.

Все перечисленные типы адвербиальных конструкций находятся на пересечении двух активных процессов, происходящих в языке, – грамматикализации и фразеологизации. Это «застывшие», устойчивые образования, для которых «наречная» (падежная или предложно-падежная) форма является «начальной», исходной.

Возникает вопрос словарного представления таких обстоятельственных оборотов. Поскольку они являются особыми единицами, то теоретически они и в словарях должны быть представлены как отдельные единицы. Однако традиционные словари – как толковые, так и фразеологические – не очень хорошо приспособлены для фиксации таких оборотов и не позволяют отразить их специфику.

ЗАМЕЧАНИЕ

В данной работе мы рассматриваем только обстоятельственные обороты, однако подобные же проблемы словарного представления и поиска возникают и в связи с несогласованными определениями типа *юбка* ***в клетку*** (***в крупную клетку***), *мальчики* ***высокого роста***, и в связи с предикативами, которые имеют форму предложно-падежных оборотов: (*не*) *по сердцу*; (*не*) *по душе*; *не в себе*; *не с руки*; *не по плечу*; *в моде*; *в ходу* и т.п.

Обычные толковые словари можно назвать «номинативными»: словарный вход представляет собой начальную форму – номинатив для имен и инфинитив для глаголов (это относится, разумеется, к словам, имеющим парадигму). Другой их особенностью, которая имеет важные последствия для наречной и служебной лексики, является «одноэлементность»: отдельный словарный вход имеют только те единицы, которые пишутся в одно слово. Предложные обороты (а также другие обстоятельственные конструкции – например, в творительном падеже, ср. *первым делом*) вступают в противоречие с «номинативным» и «одноэлементным» принципами толковых словарей и потому отражаются в них непоследовательно и противоречиво, а иногда и вовсе не получают отражения.

Наименьшую лексикографическую проблему представляют падежные формы существительных (*летом, боком, верхом* и т.п.), превратившиеся в наречия: в толковые словари они записываются как наречия и получают толкование. Сюда же нужно отнести предложные формы, которые по правилам орфографии пишутся слитно (*внизу, вверх, вмиг* и т.д.) и которые также трактуются как наречия. Сложнее обстоит дело с предложно-падежными оборотами. Если единицы, исторически восходящие к предложно-падежным сочетаниям, пишутся раздельно, они не могут быть входом в толковом словаре (хотя являются входами в орфографических словарях, т.е. признаются самостоятельными единицами). Например, в отличие от наречия *боком*, которое имеет отдельную словарную статью, наречие *бок о бок* дается в статье существительного *бок* в качестве фразеологизма. Чтобы отразить наречие *в обнимку*, приходится помещать в словарь несуществующее слово *обнимка*, естественно, без толкования. В словаре можно обнаружить еще более несуществующее слово *скак* (в угловых скобках), которое служит входом для наречия *на скаку*.

Такие же проблемы возникают с производными отыменными предлогами. Процесс превращения сочетаний абстрактных существительных с простыми предлогами в производные именные предлоги в современном языке является активным, еще не завершившимся, и список именных предлогов постоянно пополняется. Так, в словаре [6] список производных предлогов намного шире, чем в АГ-80 [1]. Например, сюда включены такие выражения, как *под прикрытием чего, под сень чего, под сенью чего, на основе чего, под фирмой чего, под флагом чего, под эгидой чего, в ранге кого, в разрезе чего, в преддверии чего, в процессе чего, в лице кого*; *во главе чего*, *по принципу чего* и др., которых не было в списке производных предлогов, приведенном в АГ-80.

При этом, если одноэлементные предлоги помещаются в словарь наравне с другими неизменяемыми словами – наречиями, частицами, союзами, – то неодноэлементные производные предлоги, как и наречия типа *бок о бок*, не включаются в обычные толковые словари в качестве отдельного словарного входа, а помещаются в

045-Kustova_G:_Layout 1 18.05.2008 15:43 Page 300

словарные статьи соответствующих существительных в заромбовой части среди фразеологизмов. Так, например, предлог *ввиду*, который пишется слитно, является отдельным словарным входом, а предлог *в виде*, который пишется раздельно, включается во фразеологическую зону существительного *вид*, т.е. приравнивается к фразеологизмам. С грамматической точки зрения это выглядит довольно непоследовательно, т.к. производные предлоги, в отличие от настоящих фразеологизмов, не употребляются как отдельные члены предложения, а присоединяют падежные формы: *в виде исключения*, *в случае дождя*, *на случай опоздания*, *по случаю праздника* и т.д. С другой стороны, на базе производных предлогов иногда возникают настоящие фразеологизмы, ср. *в случае чего*, – они и должны помещаться во фразеологическую зону словарной статьи.

То же относится к обстоятельственным оборотам с прилагательными типа *во всяком случае*. Они (несистемно и неполно) отражаются в толковых словарях во фразеологической зоне словарной статьи входящего в их состав существительного или прилагательного.

Что касается фразеологических словарей, то они устроены по-разному. Иногда поиск организован по номинативному принципу. Так, в словаре [8] даются фразеологизмы с определенным ключевым словом – например, *рука*: *легкая рука, правая рука, сбыть с рук, рука не поднимается, развести руками, руки не доходят, золотые руки, длинные руки* и под.; наречные сочетания типа *на скорую руку* в таком словаре практически не отражаются. Есть словари, в которых поиск фразеологизмов с существительными организован по падежному принципу (на каждую падежную форму отдельно), например, *дела*: *нет дела*; *деле*: *в самом деле*, *на деле* и др.; *дело*: *в чем дело*, *гиблое дело*, *дело в шляпе*, *дело с концом* и др.; *делом*: *первым делом*, *между делом* и др.; *делу*: *ближе к делу*, *делу конец* и др.). В таких словарях учитываются наречные фразеологизмы типа *на скорую руку*.

Производные предлоги отражаются во фразеологических словарях непоследовательно: например, в [9] *по случаю* есть, а *по причине* и *в результате* нет. При этом многие фразеологические словари не содержат морфологической информации и не различают, так сказать, «лексических» фразеологизмов и «грамматических» фразеологизмов. Например, в [9] производный предлог *во имя* (без грамматической пометы «предлог») дается в одном ряду с «настоящим» фразеологизмом *имя им легион*.

Кроме того, подача материала во фразеологических словарях осуществляется по знаменательным элементам, т.е. поиск по предлогу не предусмотрен (например, в [9] фразеологизм *на скорую руку* будет иметь вход *руку*). Между тем пользователь может быть заинтересован в том, чтобы найти, например, все производные предлоги с простым предлогом *в* (*в результате*, *в ходе*, *во имя*, *в противовес* и т.д.) или все обстоятельственные обороты с предлогом *на* (*на всякий случай*; *на худой конец*; *на полную катушку*; *на первых порах*; *на скорую руку*; *на живую нитку*; *на широкую ногу* и под.).

С другой стороны, существуют разные типы «неноминативных» словарей, где рассматриваемые в данной работе явления получают пусть не исчерпывающее, но более полное и системное отражение.

Во-первых, это словари служебных слов. Например, словарь [7] включает в себя: производные предлоги и союзы (*за исключением, за пределы, не в пример, по прошествии, по причине того что, в силу того что*); вводные слова (*как видишь; значит; между прочим*); грамматические обороты, выступающие в функции частиц или местоимений (*знай себе*, *неизвестно что*); фиксирует грамматические функции местоимений (*как бы ни* – союзное слово) и прономинализацию прилагательных (*последний* в знач. местоимения).

Кроме того, потребность отразить неоднословные «неноминативные» единицы выразилась в создании словарей эквивалентов слова. В первую очередь нужно упомянуть известный «Словарь эквивалентов слова» Р.П. Рогожниковой [5]. В него вошли сочетания, которые, будучи «грамматическими» или грамматикализованными единицами, обычно не попадают во фразеологические словари: это сложные союзы (*все равно что*, *даром что* и под.), производные предлоги (*в силу*, *в случае* и под.), наречные обороты (*по привычке*, *до отказа*, *в принципе* и под.), а также трехэлементные обороты типа *в конце концов*, *в случае чего*, *тот или иной*. Впрочем, трехэлементные сочетания, в силу их меньшей грамматикализованности, представлены в словаре Р.П. Рогожниковой далеко не полно: есть оборот *в какой-то мере* (в качестве наречия), но нет оборотов *в большой мере*, *в известной степени*, *в определенной степени*, *в высшей степени* (хотя их тоже можно считать аналогами слов – *отчасти*, *очень* и под.), есть наречие *в итоге*, но нет оборота *в конечном итоге*, есть оборот *в конце концов*, но нет оборота *в конечном счете*, есть обороты *в крайнем случае* и *в таком случае*, но нет *в лучшем случае*, *в противном случае*, *на всякий случай*.

Недавно вышедший словарь [6] уже вполне отчетливо отражает идею необходимости собрать в одном месте связанные между собой явления – неизменяемые слова (наречия, предлоги, союзы, частицы); обстоятельственные обороты в наречной функции, в основном, фразеологического или устойчивого характера (*смеха ради*; *с выражением*; *с глазу на глаз*; *с грехом пополам*; *по полной программе*; *из первых рук*; *до потери пульса*; *в двух словах* и т.п.); предложные обороты, выступающие в качестве несогласованных определений (ср. *в клетку*) или в качестве предикатов (*в моде*; *по плечу*; *не в себе*; *не ахти*; *не велика беда*; *цены нет*); устойчивые речевые обороты и штампы (*сдвиг по фазе*; *сделай милость*; *туши свет!*; *так нет же*; *так и так*; *так и быть*;

045-Kustova_G:_Layout 1 18.05.2008 15:43 Page 301

не тут-то было; *здравствуйте*, *я ваша тетя* и под.); вводные слова и обороты (*с вашего позволения*; *как принято говорить*; *глядишь* и под.).

Конечно, и в этом словаре имеются лакуны: например, обороты *в одну секунду* и *в один момент* есть, а оборота *в считанные секунды* нет, хотя первые нельзя считать более идиоматичными, чем последний; есть сочетания *как на беду*, *как назло*, *как на смех*, но нет сочетания *как на грех*; есть сочетание *изо всех сил*, но нет *из последних сил*; есть обороты *из вежливости*, *из принципа*, но нет *из лучших побуждений*; есть *на ваш взгляд*, но нет *на ваше усмотрение*. Это не удивительно, т.к. нет единого подхода и единых концептуальных представлений о «неноминативных» оборотах и принципах их представления в «неноминативных» словарях, как нет и теории самих неноминативных словарей. Создание же такой теории, в свою очередь, тормозится отсутствием справочников, где бы подобные языковые явления были собраны в обозримой и доступной для анализа форме.

Таким образом, кажется очевидным, что назрела и теоретическая, и практическая необходимость в создании словарей, которые позволили бы отразить специфику именно обстоятельственных и других «неноминативных» сочетаний и единиц, в которых была бы систематизирована и унифицирована подача подобных форм. Однако разработка новых типов «бумажных» словарей – вопрос времени, возможно, значительного. Другое дело – электронные словари (базы данных). Электронный словарь является наиболее оптимальным способом представления такого языкового материала, как неодноэлементные единицы и обороты. У него есть целый ряд преимуществ перед бумажным словарем: он может быть создан в гораздо более короткие сроки, у него практически нет ограничений на объем, он может оперативно пополняться и обновляться. Такой словарь может дать представление обо всем объеме подобных явлений в современном русском языке. В электронном словаре эффективно могут быть решены не только проблемы полноты, но и проблемы поиска: единицы, помещаемые в словарь, будут сгруппированы по разделам в соответствии с разными моделями, о которых говорилось выше, и снабжены разными видами лингвистически релевантной информации. Это позволит осуществлять поиск по разным группам единиц и по разным параметрам.

Вот некоторые типы возможных запросов пользователя:

поиск по компоненту:

все обороты с заданным существительным – здесь встретятся все обстоятельственные обороты «прилагательное + существительное», например: *случай*: *в любом случае*; *в крайнем случае*; *во всяком случае*; *в таком случае*; *в лучшем случае*; *на крайний случай*; *на всякий* (*пожарный*) *случай*; а также обороты других моделей – *от случая к случаю*; *по случаю* (наречие и предлог); *на случай* (предлог) и т.п.;

все обороты с заданным прилагательным – *любой*: *в любой момент*, *в любом случае*; *любыми путями*; *любой ценой*; *под любым предлогом* и т.п.; *первый*: *первым делом*; *в первую очередь*; *на первый взгляд*; *на первых порах*; *из первых рук* и т.п.;

поиск по грамматической характеристике – именные предлоги: *в адрес кого*, *в виде чего, в границах чего, в области чего, в пользу кого-чего*, *в порядке чего, в результате чего*, *в роли кого*, *в силу чего*, *в ходе чего, под видом кого-чего*, *под предлогом чего* и т.д.; обороты в творительном падеже: *первым делом, ровным счетом, полным ходом, таким образом, задним числом* и т.п.;

поиск по синтаксической функции: *в моде*, *не по душе* – предикатив, *по недоразумению*, *своими силами*, *с какой стати*, *тютелька в тютельку* – обстоятельство;

данные о том, соответствует ли обороту «пара» в виде именного предлога (*в крайнем случае – в случае* (*задержки*));

поиск по семантической функции (помете): *действительно – в действительности – на самом деле* – модальная лексика; *в любом случае* (‘что бы ни случилось’) – уступка; *изо всех сил*, *в высшей степени* – степень.

Чтобы получить обороты одного определенного типа из бумажного словаря, придется пролистать весь словарь; если же искать их с помощью поисковой машины и с привлечением Интернет-ресурсов, на выходе получим много шума.

Большим преимуществом обсуждаемой базы данных перед бумажными словарями является то, что в нее могут быть включены не только устойчивые, полностью фразеологизованные обороты, но и обороты с «альтернативным наполнением», например, *в обязательном / добровольном / приказном порядке*; *в первоочередном / плановом / пожарном / спешном / срочном порядке*; *в досудебном / судебном / законном / законодательном / порядке*; *в электронном / письменном / печатном / рукописном / бумажном виде*; *в сжатом / развернутом / сокращенном виде*; *в трезвом / пьяном виде*; *в горячем / холодном виде*; *в сыром / вареном / копченом / свежем / жидком виде*; *в сложенном / развернутом виде* и т.п. Такие сочетания не считаются фразеологизмами, тем не менее они, как уже отмечалось выше, не имеют «номинативной формы» (*в копченом виде – *копченый вид*), поэтому их тоже стоит помещать в неноминативные базы данных. Конечно, возникает вопрос, какие именно сочетания (с какими прилагательными) нужно включать в базу (коль скоро эти прилагательные альтернативны). Однако этот вопрос легко решить с помощью Национального корпуса русского языка, а именно: можно включить в базу те сочетания, которые реально встречаются в Корпусе.

045-Kustova_G:_Layout 1 18.05.2008 15:43 Page 302

Список литературы

1. АГ-80 – Русская грамматика. Т. 1. М.: 1980.
2. Кустова Г.И. Обстоятельственные группы типа во всяком случае в современном русском языке // Активные процессы в современном русском языке. Хельсинки: 2008а.
3. Кустова Г.И. Обстоятельственные конструкции с прилагательными // НТИ. Сер. 2. 2008б, № 4.
4. МАС – Словарь русского языка. В 4-х тт. Под ред. А.П. Евгеньевой. М.: 1981–1984.
5. Рогожникова Р.П. Словарь эквивалентов слова. М.: 1991.
6. Словарь наречий и служебных слов русского языка. Сост. В.В. Бурцева. М.: 2005.
7. Словарь структурных слов русского языка. Под ред. В.В. Морковкина. М.: 1997.
8. А.М. Мелерович, В.М. Мокиенко. Фразеологизмы в русской речи. Словарь. Изд. 2-е, стереотип. М.: 2001.
9. Фразеологический словарь русского языка. Под ред. А.И. Молоткова. Изд. 5-е, стереотип. СПб: 1994.

046-Lande:_Layout 1 18.05.2008 15:43 Page 303

ВЕБ-ПРОСТРАНСТВО И МАТЕРИАЛЫ ИНФОРМАЦИОННЫХ АГЕНТСТВ

WEB-SPACE AND MATERIALS OF NEWS AGENCIES

Ландэ Д.В. *(dwl@visti.net),* ***Брайчевский С.М****. (smb@visti.net),* ***Дармохвал А.Т.*** *(hval@visti.net),*
Морозов А.Ю. *(alex@visti.net)*
Информационный центр «ЭЛВИСТИ», Киев, Украина

Исследуется в какой мере материалы, доступные подписчикам информационных агентств за плату, публикуются в открытом доступе на информационных веб-сайтах. Получено распределение сообщений информационных агентств по времени запаздывания, определено как удельное количество перепечаток материалов информационных агентств на веб-сайтах, так и сообщений из Интернет, включенных в состав лент информационных агентств.

Одним из ключевых аспектов развития современных информационных технологий является специфика взаимоотношений между информационными агентствами (ИА), традиционно играющими роль поставщиков информации, и СМИ, являющимися основным ее потребителем. На взгляд авторов предлагаемой статьи, эти взаимоотношения в значительной мере устарели и нуждаются в серьезных коррективах как в технологическом плане, так и в плане организационном, включая законодательное регулирование. Главная причина такого положения дел состоит в быстром расширении влияния на информационные процессы сетевых технологий и, разумеется, в первую очередь Интернет. Развитие этих технологий привело к качественным изменениям в структуре всего процесса информирования общественности на всех его звеньях, в результате чего ситуация требует кардинального пересмотра основных механизмов, лежащих в основе функционирования медийных средств.

Информационные агентства снабжают своих подписчиков информацией на условиях, которые на сегодняшний день выглядят по меньшей мере странно. В частности, типичным условием относительно использования материалов ИА является запрет на размножение и распространение их любыми способами. Таким образом агентства пытаются защитить свою продукцию от копирования, зачастую ссылаясь на законодательство об авторских правах. Вместе с тем в статье 8 Закона РФ «Об авторском праве и смежных правах» говорится о том, что «сообщения о событиях и фактах, имеющие информационный характер» не охраняются авторским правом. В аналогичном Законе Украины в статье 10 также предусмотрено, что сообщения о новостях или текущих событиях не охраняются авторским правом. Таким образом, условия, декларируемые большинством ИА со ссылкой на законодательство об авторских правах, являются неправомерными, по крайней мере, по отношению к их основной продукции – информационным сообщениям.

Не лучше обстоит дело и с содержательным аспектом проблемы. Никто, безусловно, не ставит под сомнение авторские права на те материалы, которые действительно имеют автора в обычном смысле слова (интервью, аналитические разработки, эксклюзивные репортажи и т. д.). Но говорить об авторских правах на сообщения об официальном визите главы государства или вступлении в силу нового закона явно лишено конкретного смысла. Мы уже не говорим о текстах законов, указов и т. п., для которых законодательно предусмотрен порядок обнародования.

Как всегда в подобных ситуациях, новые тенденции начинают прокладывать себе дорогу, не дожидаясь официальных решений, что неизбежно приводит к перераспределению не только ресурсов, но и функциональных ролей участников коммуникации. Поэтому для выработки обоснованных рекомендаций желательно было бы вначале разобраться в том, что и как происходит в действительности.

Целью данной работы было исследование того, в какой мере материалы, доступные платным подписчикам основных ИА, становятся доступными в открытом доступе на информационных веб-сайтах. Ценность информационных сообщений во многом определяется оперативностью, поэтому отдельной задачей была оценка запаздывания публикаций в Интернет по сравнению с временем рассылки соответствующих сообщений. Забегая вперед, скажем, что почти в третьей части рассматриваемых случаев время задержки

046-Lande:_Layout 1 18.05.2008 15:43 Page 304

оказалось отрицательным, т.е. ИА копировали сообщения с веб-сайтов, да еще и со значительным запаздыванием.

При проведении исследований авторы получили уникальную возможность доступа к подписным материалам ведущих ИА, представленных в украинском информационном пространстве. Кроме того, в распоряжении авторов находилась система контент-мониторинга InfoStream [1] – поисковая система, с помощью которой в реальном масштабе времени сканируется свыше 3000 информационных веб-сайтов, представленных в украинском и российском сегментах веб-пространства. Таким образом, в ходе исследования рассматривались два текстовых корпуса (точнее, набора «словесных сигнатур» текстов [2], представленных в этих корпусах) – сообщений ИА и текстов, сканированных из веб-пространства. Рассматривались сообщения ИА по общеполитической тематике, поступающие в течение 5-25 ноября 2007 года. Их объем оказался равным 8955 документов. Эти сообщения сравнивались с текстами, сканируемыми из Интернет в течение всего ноября 2007 года, количество которых составило свыше 1 млн. документов.

Технически задача нахождения дубликатов (в данном случае речь идет именно о дубликатах, а не о сообщениях по той же теме, но другими словами, т. е. учитывались перепечатки с незначительными искажениями) решалась методом, который описан в [2]. Этот метод относится к группе методов нахождения «подобных» документов [3-5], основанных на выделении некоторого множества опорных слов, имеющих наибольший TFIDF [2, 3]. В качестве некоторых «инвариантов» для отдельных сообщений использовались цепочки из 12 опорных слов, прошедших процедуру морфологической обработки (стемминга). Такое небольшое количество термов в цепочке, которая является своеобразной словесной сигнатурой, объясняется небольшой средней длиной новостных сообщений (2000-3000 символов).

В результате проведенных исследований удалось получить такие данные:

- из 8955 сообщений ИА на веб-сайтах было опубликовано 5567 сообщений (62 %);
- общее количество перепечаток на различных веб-сайтах составило 39901 (456 %). Соответствующее распределение, оказавшееся гиперболическим, приведено на рис. 1;
- количество перепечаток с положительным временем запаздывания (из материалов ИА - на веб-сайты) составило 28933 (73 %);
- количество перепечаток с отрицательным временем запаздывания (перепечаток из Интернет, помещаемых в ленты ИА) составило 10968 (27 %).

Ранжированный график распределения сообщений ИА по времени задержки публикаций приведен на рис. 2, на котором четко видны экстремальные отклонения в начальной и конечной области. Отклонение в начальной области характеризует большое время задержки включения в ленты ИА материалов, размещенных, как правило, на сайтах органов государственной власти (инертность ИА, отсутствие у них средств мониторинга веб-пространства). Отклонения в конечной области объясняются задержками перепечаток на веб-сайтах сообщений, получивших со временем некоторое новое продолжение. Вместе с тем центральная область графика (от 1000-го по 5000-е сообщение) имеет стабильный характер со средним значением около получаса.

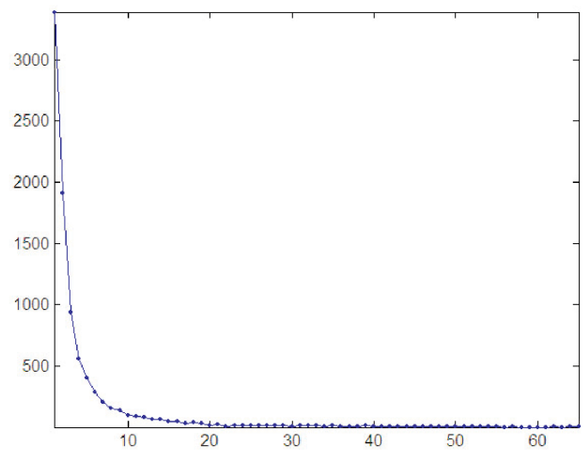


Рис. 1. Количество сообщений ИА (ось ординат), ранжированное по количеству перепечаток на веб-сайтах (ось абсцисс). Значению 0 соответствует количество неперепечатанных сообщений

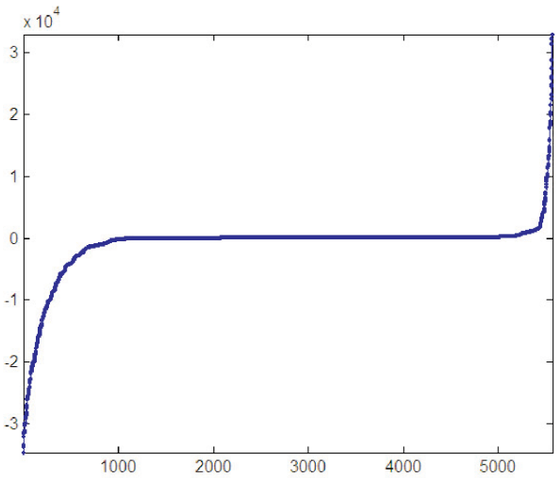


Рис. 2. Распределение сообщений ИА (ось абсцисс) по времени запаздывания в минутах (ось ординат)

046-Lande:_Layout 1 18.05.2008 15:43 Page 305

Массовый характер перепечаток позволяет делать выводы о том, что все сообщения, интересные администраторам соответствующих веб-сайтов, были перепечатаны. По-видимому, примерно 38 % сообщений ИА оказались им недостаточно интересными.

В заключение можно сделать несколько выводов. С точки зрения технологий оказалось, что методы определения нечетких дубликатов сообщений, развитые в последние годы как отечественными, так и зарубежными исследователями, оказались очень интересными в рассматриваемом применении. Результаты заставляют задуматься, за что же платят подписчики информационным агентствам сегодня, когда большая часть информации с минимальной задержкой доступна в Интернет, а полноту могут обеспечить системы контент-мониторинга? По-видимому, за аналитический подбор этой информации, репрезентативность и достоверность. То есть, информационное агентство, если оно желает выжить в современных условиях, должно уделять повышенное внимание именно аналитической обработке информации, превращаясь в агентство информационно-аналитическое.

Естественно, полученные результаты могут учитываться также разработчиками информационно-поисковых систем и систем контент-мониторинга. Усиление аналитической составляющей таких систем уже сегодня позволяет им выступать на рынке рядом с крупнейшими информационными агентствами.

Список литературы

1. Григорьев А.Н., Ландэ Д.В. и др. Мониторинг новостей из Интернет: технология, система, сервис: научно-методическое пособие. – К.: ООО «Старт-98», 2007. – 40 с. (http://dwl.visti.net/art/booklet/ booklet.pdf)

2. Д.В. Ландэ, А.Т. Дармохвал, А.Ю. Морозов. Подход к выявлению дублирования сообщений в новостных информационных потоках // Труды 8ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2006, Суздаль, Россия, 2006. – С. 115-119. (http://www.rcdl2006.uniyar.ac.ru/papers/ paper_71_v2.pdf)

3. Ю.Г. Зеленков, И.В. Сегалович Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Переславль, Россия, 2007. – Том 1, С. 166-174. (http://rcdl2007.pereslavl.ru/papers/paper_65_v1.pdf)

4. Никконен А.Ю. Устранение избыточности и дублирования сюжетов новостных сообщений // Интернет-Математика. Сборник работ участников конкурса. – Екатеринбург: Изд-во Урал. Ун-та, 2007. –С. 157-167 (http://download.yandex.ru/IMAT2007/imat2007.pdf)

5. J. Bourdaillet. Alignment of Noisy Unstructured Text Data // IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data. Hyderabad, India - January 8, 2007. P. 139-146 (http://research.ihost.com/and2007/ cd/ Proceedings_files/p139.pdf)

047-Levontina_I:_Layout 1 18.05.2008 15:44 Page 306

ЗАГАДКИ ЧАСТИЦЫ *УЖ*[1]

THE RIDDLES OF THE RUSSIAN PARTICLE *UZH*

Левонтина И.Б. *(irina.levontina@mail.ru)*
Институт русского языка им. В.В.Виноградова РАН

Русская дискурсивная частица уж (Нет уж, Сиди уж, Уж простите) трудна для описания, поскольку неясно, каким образом создаваемые ею прагматические эффекты связаны с компонентами ее значения. В работе делается попытка выделить некоторые компоненты значения частицы и установить такие связи.

Модальная частица *уж* – действительно одно из самых загадочных слов русского языка. С одной стороны, она чрезвычайно выразительна: такие яркие фразы, как *Да уж* и *Нет уж*, *Ты уж прости* и *Ладно уж*, *Где уж мне* и *Уж я-то знаю* создают впечатление, что у частицы *уж* должно быть очень отчетливое значение. С другой стороны, оказывается невероятно трудно построить такое описание *уж*, которое охватило бы все эти и бесчисленные другие типы контекстов. Не случайно Е. В. Урысон описывает модальное *уж* как «особое разговорное словечко с почти неуловимой семантикой, некий «наполнитель» высказывания, придающий ему идиоматичную разговорную окраску» [Урысон 2007: 539].

Нельзя сказать, что *уж* было обойдено вниманием исследователей. Эту частицу описывали, ей даже специально посвящено некоторое количество работ [Paillard 1986-87, Mendoza 1999, 2000, Урысон 2007]. Довольно подробное перечисление контекстов употреблений частицы *уж* содержится в таких словарях, как «Словарь структурных слов русского языка» (ред. В. В. Морковкин) [Морковкин 1997], «Словарь русских частиц» Э. Шимчук и М. Щур [Шимчук, Щур 1999]. Однако, как часто бывает с подобными словами, словарные описания сводятся к списку никак друг с другом не связанных употреблений, причем семантический вклад частицы в каждом случае остается неясным. Другая крайность состоит в поиске инварианта значения, который оказывается в результате настолько абстрактным, что на его основе нельзя объяснить смысловые нюансы конкретных словоупотреблений, а уж тем более – предсказать их; ср. [Paillard 1986-87, Mendoza 1999, 2000].

Недостатки обоих крайностей хорошо видны на примере интерпретации основного круга контекстов *уж*, который в словарях обычно рассматривается как реализация первого из его модальных значений. Этот круг представлен фразами типа: *Уж за это я ручаюсь!, Уж я-то знаю; Знает и географию, и математику, а уж о литературе и говорить нечего!; Уж кто-кто, а я могу это сказать!* В словаре [Шимчук, Щур 1999] *уж* здесь толкуется в том смысле, что оно «выделяет из некоторого множества предмет, признак или событие, которые в ситуации выбора по тем или иным причинам должны быть рассмотрены в первую очередь».

Обратим, однако, внимание на следующее обстоятельство. Во всех примерах этого класса *уж* в теме, причем часто это тема, маркированная при помощи *–то (Уж он-то)* или конструкции типа *уж кто-кто* [Левонтина 2003]. Значительная часть смысла связана с этим. Так где же смысл *уж*?

Сопоставим две фразы: *Уж я-то это знаю* vs. *Я-то это знаю*. Разница между ними довольно очевидна. Фраза без *уж* просто противопоставляет один из элементов множества другим его элементам, тогда как фраза с *уж* вводит некую шкалу, относительно которой элементы множества упорядочены таким образом, что один из них занимает на шкале крайнее положение и обладает каким-то свойством в большей степени, чем можно было ожидать.

[1] Данная работа выполнена при поддержке следующих грантов: гранта Президента РФ на поддержку ведущих научных школ № НШ-56.11.2006.6, гранта РГНФ № 06-04-00289а на «Разработку словника и проспекта активного словаря русского языка», гранта Программы фундаментальных исследований ОИФН РАН «Русская культура в мировой истории».

[2] Временное *уж* как вариант или синоним *уже* здесь не рассматривается.

047-Levontina_I:_Layout 1 18.05.2008 15:44 Page 307

Идея превышения ожиданий[3]:

Из-за идеи превышения ожиданий оказываются невозможными фразы:

(1)**Уж я-то это хоть немного знаю;*

(2) ??*Уж математику-то я, может быть, сдам.*

Идея шкалы:

Можно сказать:

(3) *Я-то это знаю, а вот другим невдомек*,

но *уж* было бы в этом случае неуместно. *Уж я-то это знаю* предполагает, что другие, возможно, знают тоже, но обо мне это можно сказать совершенно точно. Именно эта идея шкалы и полюса и есть, в сочетании с идеей превышения ожиданий, собственный вклад частицы *уж* в значение высказывания, в то время как идея противопоставления, возможно, задается контекстом. Более того, *уж* даже отчасти гасит противопоставление, поскольку оно показывает, что остальные элементы множества не исключаются из рассмотрения, а просто данный элемент оказывается особо выделен.

Наличие в значении модального *уж* компонентов ‘шкала’ и ‘превышение ожиданий’ совершенно естественно: они связаны с тем временным значением *уж*, в котором оно близко к временному *уже* и указывает на опережение. Развитие у временного слова модального значения, при котором происходит трансформация временной шкалы в шкалу ожиданий, вполне типично. Об этом применительно к значению частиц *уже* и *уж* см. [Урысон 2007; 536-360].

Между тем, толкование рассмотренных контекстов с контрастной темой лежит в основе того инвариантного значения, которое приписывается *уж* в тех работах, где описание выполнено в идеологии инвариантов. Такие толкования могут казаться правдоподобными и узнаваемыми, поскольку они хорошо ассоциируются с большим кругом контекстов *уж*, однако по сути они не решают загадку этой частицы. Примером здесь может служить работа Д. Пайара «*Уж,* или необсуждаемое»[Paillard 1986-87]. Смысл его толкования таков: *уж* маркирует некоторую ценность p как не подлежащую обсуждению и одновременно противопоставляет ее ценности p’, которая эксплицитно или имплицитно задана в предшествующем контексте, однако в качестве преодоленной, отвергаемой, недействительной.

Сходно с пайаровским толкование Цыбатова [Zybatow 1990]: *Уж* указывает на выраженную в левом контексте или выводимую неуверенность в том, что входящее в сферу действия частицы утверждение соответствует действительности и подчеркивает, что оно не может не соответствовать действительности. На эту же тему – толкование И. Мендосы [Mendoza 2000]:

уж

(a) маркирует какую-то ценность W как действительную;

(b) причем имплицируется также ценность W’, которая недействительна;

(c) причем ценность W’ имплицитно или эксплицитно фигурирует с предшествующем тексте или может быть выведена.

Это толкование также ориентировано на первый круг контекстов (*Уж я-то знаю*, *уж кто-кто*). Однако если взять фразы *Ладно уж, Ты уж меня прости, Я уж лучше подожду, Не знаю уж, почему* и подобные, на них оно натягивается с трудом.

Лучше рассмотреть отдельные типы контекстов, выяснить, какой вклад вносит в них *уж*, а потом думать, можно ли все свести воедино или здесь разные единицы. Конечно, в небольшой статье охватить все контексты употребления *уж* невозможно, но рассмотрим хотя бы некоторые из них. И хотя высказывания типа *Уж кто-кто, Уж мне ли этого не знать, Уж Петя-то не подведет* и подобные весьма яркие, как раз в них семантический вклад *уж* совершенно не является определяющим. Во всех этих случаях *уж* можно опустить, и потеря смысла не будет критической, поэтому они для нас не так существенны.

Здесь мы рассмотрим более подробно употребление *уж* в побудительных высказываниях. Этот выбор не случаен: в побудительных высказываниях говорящий должен заботиться о мотивировках, тщательно выстраивать свои отношения с адресатом, поэтому, с одной стороны, они богаты разнообразными прагматическими эффектами, а с другой – значения частиц в них относительно менее аморфны.

[3] Надо сказать, что идея превышения ожиданий представлена в самых разных употреблениях уж:

Уж это-то я знаю! (о превышении ожиданий адресата);

Что-то уж очень много (превышены ожидания говорящего);

Ладно уж. (Превышены усилия, на которые говорящий имел в виду пойти);

Нет уж! Ну уж нет! (Говорящий хочет вдребезги разбить ожидания адресата, который ожидал легко получить согласие); Однако и *Да уж!, Уж что да, то да!* (Звучит иначе, чем *Да, конечно.*) Ожидание состоит в том, что собеседник думает, что меня надо в этом убеждать. А я не присоединяюсь к его мнению, а высказываю свое, еще более радикальное. Возможно, впрочем, иронически.

047-Levontina_I:_Layout 1 18.05.2008 15:44 Page 308

Итак, если приложить приведенные выше толкования к побудительному высказыванию, должно получиться что-то вроде:

(Говорящий побуждает адресата сделать что-то)

+ *уж* – говорящий вводит в рассмотрение возможность невыполнения этого, однако отвергает эту возможность, а выполнение адресатом действия рассматривает как не подлежащее обсуждению.

Между тем, если посмотреть на совершенно типичную фразу с *уж*: *Ты уж не спорь с ним*, можно увидеть, что здесь имеет место нечто почти противоположное; говорящий признает, что адресат имеет право не принимать его просьбу во внимание, но все же просит выполнить ее.

Ср. характерный пример:

(4) *Пройдя комнату и нерешительно подойдя ко мне сбоку, ..., мать, вынув руку из-за пазухи, положила мне на стол две смятых, словно желающих стыдливо уменьшиться, пятирублевых бумажки. Погладив затем своей скрюченной ручкой мою руку, она тихо сказала: - Ты уж прости меня, мой мальчик* (М. Агеев, Роман с кокаином).

Здесь выражается полная униженность и никакой нет никакой речи об уверенности или представлении о выполнении просьбы как о чем-то само собой разумеющемся. Подобное значение выражается при помощи совсем другой частицы – *давай* (*Давай сходи за хлебом*) [Левонтина 2005].

В словаре [Шимчук, Щур 1999] отмечается, что *уж* в побудительных высказываниях предполагает взгляд «снизу вверх», «указывает на признание говорящим авторитета адресата или на зависимость от адресата»:

Уж ты не показывай никому мою работу. Я не хочу, чтобы надо мной смеялись.

Уж вы помогите моему сыну! Век буду бога молить за вас!

Очевидно, однако, что этот взгляд снизу вверх – не часть значения *уж*, а прагматический эффект, который возникает в определенных условиях. При этом такой эффект возникает не всегда; ср:

(5) - *Так растрогали, так растрогали... Спасибо вам, - говорил Филипп Филиппович, - голубчик, я иногда на вас ору на операциях. Уж простите стариковскую вспыльчивость. В сущности ведь я так одинок...* (М. Булгаков, Собачье сердце);

В словаре [Морковкин 1997] у приимперативного *уж* выделяются два отдельных значение – «снизу вверх» или наоборот, «для выражения фамильярной покровительственности, добродушного превосходства, снисходительности, назидательности»:

Пусти уж, помощничек! Кто так иголку держит?

Подожди уж ты со своими советами, тут и без тебя советчиков много.

Очевидно при этом, что в приведенном выше примере из «Собачьего сердца» нет никакого «добродушного превосходства» и тем более снисходительности, назидательности».

Существенно при этом в том, что побудительные контексты с *уж* вообще не сводятся к этим двум типам, имеется еще много модификаций.

Часто *уж* появляется в контексте оправдания, которое может и вводиться эксплицитно:

(6) *Эллен, ты угощай завтраком гостя, а я займусь пьесой... Уж извините меня... Завтра утром сдавать надо... Посидите с женой* (В. Гиляровский. Москва и москвичи)[4].

Здесь нет ни униженности, ни превосходства, а есть желание примирить два возможные точки зрения на ситуацию: с одной стороны, нехорошо, когда хозяин бросает гостей, но с другой стороны, в данном случае есть уважительная причина. По этой же причине часто используется при выражении отказа: *Уж простите, но это невозможно.*

Интересно, что идея объяснения, ссылки на важные причины возникает не только в императивных контекстах, но и, например, фразы типа *Уж очень красивая, Уж очень толстый* неуместны при простом описании. Они возникают тогда, когда говорящий хочет обосновать свою оценку данного человека, сослаться на то, что она никак не могла быть иной; ср. об «оправдывающейся» тональности *уж* [Урысон 2007; 539].

Ср. также следующий пример. Естественно:

(7) *Что-то сковорода уж очень горячая – это так надо?*

Но неправильно или, вернее, прагматически трудно представимо:

(8) *?Сковорода должна быть уж очень горячей.*

Фраза кажется странной, но на самом деле можно себе представить ситуацию, при которой она допустима: например, мать объясняет, почему боится разрешить дочери готовить какое-то блюдо.

[4] Ср. аналогичные примеры: - До свидания. - То есть как это до свидания? - опешил администратор. - Вы уж простите, дочура у меня там спит (В. Аксенов, Пора, мой друг, пора); - Маму жалко, Витя, - сказал он басом. — Ты уж постарайся все это... сгладить как-то. Я молчал (В. Аксенов. Звездный билет)

047-Levontina_I:_Layout 1 18.05.2008 15:44 Page 309

Во многих случаях *уж* показывает, что говорящий предлагает считать данное объяснение достаточным. Сравним две фразы: *(- Почему?) – Так здесь принято* или *Так уж здесь принято.* Фраза без *уж* допускает продолжение дискуссии, в то время как фраза с *уж* призывает удовлетвориться таким ответом. Особенно показательно, что *уж* может использоваться для симуляции объяснения; ср.: - *Почему? – Так уж <Да уж так>; Уж я знаю.*

Возвращаясь к императивным контекстам *уж*, отметим, что для этой частицы очень характерно использование в ситуации, когда просят за другого; (например, отец просит за дочь:

(9) *Разреши уж ей сегодня надеть новое платье!*

Это не случайно. Отец понимает, почему мать против, он даже согласен с ней, но понимает и дочь.

По своей «примирительной» тональности подобные фразы с *уж* очень похожи на высказывания, выражающие неохотное согласие; ср:

(10) *Ладно, надень уж сегодня новое платье!*

[имеются, конечно, соображения против, они не отменяются, но можно сделать исключение].

Вообще идея «примирения» с отклоняющейся от ожиданий действительностью прослеживается в самых разных фразах с *уж*; ср.:

(11) *Такой уж я человек* [и ничего тут не поделаешь];

(12) *Это уж как водится* [не стоило и надеяться на другое];

(13) *Так уж получилось* [поздно спорить].

Уж – не единственная частица, которая используется при императиве для усиления иллокутивного эффекта. Так могут использоваться и такие частицы, как *ну, да, же*. Но все они реализуют разные «стратегии уговаривания» (см. [Левонтина 1999]).

Чем, например, отличается *Уж разреши* от *Ну разреши*, *Уж прости* от *Ну прости*? И там и там говорящий понимает, что есть основания не разрешить, не простить. Но в случае *ну* он просто нажимает, давит, канючит, не прибегая к помощи аргументов, а в случае *уж* – сам считает, что надо простить и располагает аргументами. При этом, в отличие от фраз *Да разреши*, *Да прости,* которые подразумевают, что адресат уже отказал, но говорящий настаивает, ссылаясь на то, что причины для отказа были несущественными, в случае с *уж* ситуация иная. Говорящий сам понимает, что адресат может быть против, но ссылается на важные причины сделать все же так, как он говорит. Как же получается тогда эффект смягчения просьбы? Присутствие во фразе двух точек зрения на ситуацию, видимо, понимается как подчеркнутое ненавязывание своей позиции: прошу, не пытаясь вам навязать мнение о том, что это легко или что это нужно.

Но как же быть с раздраженными фразами типа

(14) *Ты уж завтра не опаздывай!*

Судя по всему, мы имеем здесь дело с косвенным речевым актом: категорическое требование лексически оформляется как мягкая просьба, что придает высказыванию дополнительную эмоциональность: ср. похожие фразы:

(15) *Будь так любезен, не опаздывай завтра!;*

(16) *Не сочти за труд прийти завтра вовремя!*

Немного другая ситуация имеет место в процитированном выше примере: *Пусти уж, помощничек!* Она выражает неохотное согласие либо на то, чтобы адресат не выполнял порученное, либо, скорее всего, на то, чтобы говорящий сам это выполнил. *Пусти уж* звучит здесь как *Ладно уж, пусти*.

То, что в составе императивного речевого акта *уж* создает разнообразные прагматические эффекты, вполне естественно: частица воздействует на разные компоненты семантики речевого акта (‘говорящий хочет’, ‘говорящий знает, что адресат не обязан’ или, напротив, ‘считает, что адресат должен’, ‘говорящий считает, что так будет лучше для адресата’ и т. д.). При этом важны и такие аспекты ситуации, как соотношение говорящего и адресата, трудность просьбы и т. д. (см. подробнее об этом применительно к другой частице – *ка* в [Левонтина 1991]).

Обсуждение механизма получения противоположных смыслов из одного источника заставляет вспомнить работу И. Богуславского о сферах действия частицы *уже*. Там, в частности, рассматриваются два типа примеров:

(17) *Он пришел уже в десять часов, и мы все успели обсудить до начала заседания* (рано);

(18) *Он задержался на работе и пришел уже в десять часов, когда все собирались расходиться* (поздно)[5].

В первом примере на исходную пропозицию накладывается ожидание того, что данная ситуация

[5] Ср. также:

Моцарт выступал с концертами уже в пять лет.

Шагал написал картину «Большой цирк» уже в преклонном возрасте.

047-Levontina_I:_Layout 1 18.05.2008 15:44 Page 310

произойдет позже, чем указано, а во втором – конечно не того, что десять часов наступят позже, а того, что данный момент времени, заданный в данном случае при помощи события – временного ориентира, будет характеризоваться более ранней ситуацией (например, было восемь часов).

«Эти два типа ожиданий допускают очень естественное обобщение. В обоих типах употреблений слово *уже* сообщает, что реальность опережает ожидания» [Богуславский 1996: 229].

Два свойства *уже*: возможность апеллировать к ожиданиям, соотносящимся с разными аспектами ситуации, и возможность метафорического переосмысления временной оси *(Ну, это уже никуда не годится!)* – являются, вероятно, источником многих особенностей модального *уж*, о которых шла речь выше.

Видимо, где-то здесь и источник двух основных тональностей *уж* – наступательной и примирительной. В обоих случаях имеется «зазор» между положением дел и ожиданиями, но идея зазора по-разному реализуется в разных употреблениях уж. В одних случаях говорящий выражает удивление по поводу того, что действительность превосходит его ожидания, в другом – предлагает признать, что действительность не совпадает с ожиданиями и примириться с этим.

Возможно, с этим одновременным присутствием двух точек зрения на ситуацию связана и еще одна черта модальной частицы *уж*: она часто подразумевает, что говорящий различает и даже, возможно, противопоставляет свои ожидания и ожидания адресата. Действительно, для *уж* в самых разных контекстах характерна некая отстраненность от адресата: она принимает форму либо недоверия, либо полемического задора, обиды, иронии и т. д. Благодаря этому свойству высказывания с *уж* часто насыщаются весьма сложными эмоциями.

Это яркая особенность *уж*, отличающая его от большинства других русских диалогических частиц. Им, напротив, свойственна, такая гиперкооперативность, преодоление дистанции, а не ее создание. Ср. *ишь*, *аж, а?* (см. [Левонтина 1999, 2000, 2004]).

Разумеется, приведенные соображения никак не отменяют необходимость подробного описания частицы *уж*, охватывающей все типы контекстов ее употребления и объясняющие возникновение различных прагматических эффектов. Однако, как сказано в уже цитированной работе И. Богуславского о частице *уже*, «Возводя все употребления *уже* <…> к единому семантическому источнику, мы не готовы идти столь далеко, чтобы утверждать, что перед нами во всех отношениях единая лексема. С лексикографической и практической точки зрения, возможно, целесообразно разнести эти употребления по разным лексемам, хотя бы потому, что в этом случае толкования окажутся менее абстрактными и потому более наглядными. Наше же стремление состоит в том, чтобы достичь более высокой степени обобщения» [Богуславский 1996: 254].

Список литературы

1. Богуславский И. М. Сфера действия лексических единиц. М., 1996

2. Левонтина И. Б. Словарная статья частицы -ка // Семиотика и информатика, вып. 32, Москва, 1991. (136-140).

3. Левонтина И. Б. Стратегии уговаривания: частицы в повторных просьбах // Язык. Культура, Гуманитарное знание. Научное наследие Г. О. Винокура и современность. М., 1999 (с. 188-201).

4. Левонтина И. Б. Русское аж: полисемия и синонимия // Linguistische Beiträge zur Slawistik aus Deutschland und Österreich. VII JungslawistInnen-Treffen. Tübingen-Blaubeuren 1998. Verlag Otto Sagner. München 1999.

5. Левонтина И. Б. Русское финальное а?: Портрет невидимки // Слово в тексте и в словаре. Сборник статей к семидесятилетию академика Ю. Д. Апресяна. М., 2000.

6. Левонтина И. Б. Уж кто-кто (об одной конструкции малого синтаксиса) // Русистика на пороге XXI века: проблемы и перспективы. М., 2003. (116-120)

7. Левонтина И. Б. Ишь // Сокровенные смыслы. М., 2004.

8. Левонтина И. Б. Давай-давай. Язык. Личность. Текст. Сборник статей к 70-летию Т. М. Николаевой. М., 2005.

9. Словарь структурных слов русского языка (ред. В. В. Морковкин). М., 1997.

10. Урысон Е. В., Уже и уж: вариативность, полисемия, омонимия? // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2007». М., 2007

11. Шимчук Э., Щур М.. Словарь русских частиц. Berlin, 1999.

12. Mendoza I. Uzhe und uzh in der modernen russischen Standardsprache. In: Die Welt der Slaven 44, 2: 213-224. 1999.

13. Mendoza I. Zur Geschichte von Partikeln: russisch uzhe und uzh. Linguistik online 6, 2/00

14. Paillard, D. Uz ou l'indiscutable.In:Bulletin de linguistique générale et appliquée 13, 190–213. 1986/87

15. Zybatow, L. Was die Partikeln bedeuten. Eine kontrastive Analyse Russisch–Deutsch. München. 1990

048-Leontjevy:_Layout 1 18.05.2008 15:45 Page 311

У НАС У СТАТЬИ НАЗВАНИЕ НЕ ПРИДУМАЛОСЬ: КОНСТРУКЦИИ С РЕКУРСИВНОЙ ГРУППОЙ *У + GEN* В РУССКОМ ЯЗЫКЕ[1]

U NAS U STATJI NAZVANIE NE PRIDUMALOS': RUSSIAN CONSTRUCTIONS WITH RECURSIVE NP WITH PREPOSITION *U*

Леонтьева А.Л. *(njusha-nn@mail.ru),* ***Леонтьев А.П.*** *(taonick@yandex.ru)*
Московский государственный университет им. М.В. Ломоносова

В русской разговорной речи нередко можно встретить конструкциям с рекурсивной группой «у + GEN». Нарпимер, У меня у подруги кошка как-то всё мясо из раковины стащила. В статье описываются лексико-семантические, синтаксические и дискурсивные ограничения употребления таких конструкций и делается попытка объяснения этих ограничений.

1. Объект исследования и постановка задачи

Данная работа выполнена в рамках коллективного проекта по изучению когнитивной сопряженности под руководством А. Е. Кибрика ([Кибрик и др., 2006], [Кибрик 2006], [Брыкина 2007], [Леонтьев, Леонтьева 2006], [Хитров 2005] и др.)[2]. Основным объектом изучения рабочей группы является атрибутивная синтагма (1) и синонимичные ей конструкции типа (2) – (5). Последние будем называть конструкциями с внешним посессором (КВП).

(1) ***Рука девушки*** *была чем-то измазана.*

(2) ***Рука у девушки*** *была чем-то измазана.*

(2') ***У Маши руки*** *всегда чем-нибудь измазаны*.

(3) *Иван поцеловал* ***девушке руку****.*

(4) *Иван погладил* ***девушку по руке.***

(5) *Иван схватил* ***девушку за руку.***

В данной работе речь пойдет об одной из разновидностей КВП – конструкции с рекурсивной группой *у + GEN* (см. (6))[3].

(6) ***У меня у брата друг*** *вчера из армии вернулся*.

Далее для краткости будем называть такие конструкции УУК (от *у-у*-конструкция). УУК синонимична атрибутивной синтагме с двумя посессивными отношениями: *X (Y-a (Z-a))*: *друг моего брата*, *рукав Машиной куртки, собака Васиного знакомого* и т. п.

Тот член УУК, которым управляет первый предлог *у* (*меня в* примере (6)), будем называть первым посессором, тот, которым управляет второй предлог *у* (*брат* в примере (6)), - вторым посессором, а тот, который соответствует вершинному имени синонимичной атрибутивной синтагмы (*друг* в примере (6)), - обладаемым.

Наиболее близким аналогом УУК (*у X-а у Y-а Z*) является КВП с предлогом *у*, маркирующая посессор как глобальную тему дискурса – *у Y-а Z* (ср. (2), (2')). Эта конструкция возможна практически для любого смыслового отношения между посессором (Y) и обладаемым (Z)[4].

Одно из существенных отличий УУК от стандартных КВП состоит в ее стилистической закрепленности. УУК – принадлежность почти исключительно устной разговорной речи. Это свойство создало некоторые трудности при изучении конструкции. Корпус, находящийся в распоряжении рабочей группы, созданный на основании художественных текстов конца XIX – XX века, не включает ни одного примера употребления рекурсивного *у*. Более или менее представительного электронного корпуса современных разговорных текстов на сегодняшний день, насколько нам известно, не создано. В Русском Национальном корпусе (www.ruscorpora.ru)

[1] Работа выполнена при поддержке гранта РФФИ № 05-06-80288.

[2] Авторы благодарны А. Е. Кибрику и К. Ю. Дружкину за ценные замечания, высказанные при обсуждении чернового варианта статьи. Отдельное спасибо М. М. Брыкиной, любезно согласившейся вычитать окончательный вариант. Все ошибки и неточности целиком лежат на совести авторов.

[3] Альтернативное исследование данной конструкции представлено, в частности, в [Циммерлинг 2000].

[4] Подробнее о таких конструкциях см. [Леонтьев 2008]. Частный случай у-конструкций, выполняющих дискурсивную функцию, рассматривается также в [Иорданская, Мельчук 1995].

048-Leontjevy:_Layout 1 18.05.2008 15:45 Page 312

удалось найти всего 73 примера интересующей нас конструкции – практически все они представляют собой фрагменты интервью или форумов. По результатам поиска в системе Яндекс (www.yandex.ru) было также отобрано 306 примеров употребления конструкции в электронном дискурсе (форумы, блоги и т. п.)[5].

Задача, которую мы ставим перед собой, - описать особенности употребления УУК в русском языке (этому будет посвящена вторая часть данной статьи) и объяснить существующие ограничения этого употребления (все найденные объяснения следует искать в третьей части).

2. Ограничения на употребление конструкции

2.1. Синтаксические ограничения

2.1.1. Синтаксическая позиция. Как и конструкция с одинарным *у*, УУК возможна для любой позиции обладаемого. Ср примеры (7) – (10).

(7) ***У нас у коровы теленок*** *родился!* - вершина субъект.

(8) ***У меня у сестры сына*** *вчера на улице так избили…* - вершина прямой объект.

(9) ***У меня у подруги сыну*** *прислали повестку.* - вершина непрямой объект.

(10) ***У меня у сестры в доме*** *поселилась нечистая сила* - вершина косвенный объект / сирконстант.

2.1.2. Порядок слов. В большинстве случаев (почти 92% примеров Национального корпуса) УУК употребляется в начале предложения или, по крайней мере, клаузы (см. примеры выше). Начало это или абсолютное, или «прикрытое» обращением, союзом, частицей, вводным элементом или междометием.

Заметим, что для конструкции с одинарным *у* тяготение к началу клаузы не характерно: в корпусе, используемом нашей рабочей группой, таких примеров всего 31% (включая сравнительные обороты).

Примеры употребления УУК в неначальной позиции единичны. Это, во-первых, прямой или косвенный вопрос ((11) – (13)), где, естественно, начальную позицию занимает вопросительное слово. Интересно, что тяготение УУК к началу настолько велико, что она стремится по возможности разорвать вопросительную ИГ и встать как можно ближе к началу ((12), (13)).

(11) *А* ***почему*** *у тебя у мужа такое мнение о пуделях… даже не знаю* [ЭД].

(12) *А* ***сколько*** *у тебя у героев* ***жизней****(я про последнюю битву)???* [ЭД].

(13) *А* ***какая*** *все-таки у нее$_i$ у Бони$_j$* ***профессия****?* [ЭД].

Во-вторых, УУК способно вытеснить из начальной позиции тематическое подлежащее ((14) – (15)).

(14) [Обсуждается вопрос, где достать хорошую клубную музыку – «клубнячок»]. *Ты для начала перестань электронную музыку клубнячком называть... **Клубнячки** у тебя у бабушки в огороде растут... вульгарно как-то звучит...* [ЭД].

(15) *Это* ***вакумник*** *у нее$_i$ у левой фары$_j$ вынесен что ли?* [ЭД].

В-третьих, предшествовать УУК может тематическое обстоятельство (16).

(16) *Я знаю / что в медицинском то же самое / покупают экзамены.* ***В медицинском*** *у меня у сына подруга* [ruscorpora].

Порядок слов в примерах (14) – (16) воспринимается носителями языка как инверсивный, маркированный.

2.1.3. Морфосинтаксическое единство конструкции. Прототипически конструкция употребляется в нерасщепленном виде: обе ИГ с предлогом *у* расположены контактно (16). Однако расщепить УУК отнюдь не невозможно. Группа подлежащего разрывает конструкцию легче, чем группа сказуемого ((17) – (19)).

(17) ***У меня у доченьки*** *первый зубик вылез!* [ЭД]

(18) ***У меня*** *первый зубик* ***у доченьки*** *вчера вылез!*

(19) *?****У меня*** *вылез первый зубик* ***у доченьки****!*

Эксперимент показал, что на возможность расщепления конструкции также влияет и тип семантической связи между ее компонентами, или, в других терминах, тип генитивного отношения. При прочих равных конструкция с именами родства (20) расщепляется хуже, чем, например, конструкция с собственно посессивным отношением или отношением части-целого (21). [6]

(20) ??/**У меня сестру*** *вчера в больницу увезли* ***у бабушки*** *– сегодня к ней поедем навестить.*

(21) ок/?***У меня молнию новую*** *вчера в мастерской вставили* ***у куртки*** *- от старой не отличишь.*

2.2. Ограничения на предикат. Для конструкции с одинарным *у* существуют достаточно жесткие

[5] Далее, в целях экономии места, примеры из Национального корпуса будут помечаться как «ruscorpora» (без указания конкретного автора), а примеры из собранного корпуса электронных текстов – как «ЭД» (электронный дискурс); примеры, полученные экспериментально, даются без каких-либо помет. Орфография и пунктуация авторов электронного дискурса сохранена.

[6] 14-ти испытуемым предлагалось оценить допустимость 12 высказываний, где УУК с разными генитивными отношениями разрывалась разными членами предложения. Пример (21) разрешили восемь человек, пример (20) – только один.

048-Leontjevy:_Layout 1 18.05.2008 15:45 Page 313

ограничения на предикат. Принято считать, что она возможна только (или, по крайней мере, главным образом) при стативных глаголах (ср. [Падучева 2004], [Кибрик 2000] с дальнейшей библиографией). Оставив в стороне вопрос о том, насколько данное мнение соответствует действительности, отметим, что, для УУК ограничений на предикат практически не существует (см. Таблицы 1 и 2).

тип предиката	**пример предиката**	**возможно ли?**	**пример конструкции**
действие	идти, писать	+	***У меня у подруги*** *сын диктант* ***писал****, забыл точку в последнем предложении поставить, так ему оценку на бал ниже поставили.*
деятельность	руководить, воевать	+	***У меня у подруги*** *сестра там аспирантами* ***руководит*** ***У меня у друга*** *дед на той войне* ***воевал***
занятие	играть, гулять	+	***У меня у подруги*** *дети здесь* ***играют****, а вы…*
поведения	баловаться	+	***У меня у одной знакомой*** *сын тоже со спичками* ***баловался*** *– теперь в больнице лежит.*
воздействие	размывать (дожди, полотно)	-, из-за ограничений на актанты	****У нас у реки*** *вода все берега* ***размыла***
процессы	кипеть, сохнуть	+	***У меня у бабушки*** *чайник по полчаса, бывает,* ***кипит*** *– она не слышит* ***У меня у бабушки*** *белье всегда долго* ***сохнет***
проявления	блестеть, пахнуть	+	***У меня у бабушки*** *кастрюли всегда* ***блестят*** *как новые*
состояние	болеть	+	***У меня у сына*** *живот* ***болит***
свойство	заикаться	+	***У меня у подруги*** *сын* ***заикается***
параметр	вмещать, насчитывать, весить	+	***У меня у подруги*** *сын уже двенадцать кило* ***весит.***
события	встретить (случайно)	+	***У меня у подруги*** *сын вчера Пугачеву на улице* ***встретил.***
положения	висеть		***У меня у бабушки*** *ключи под ковриком* ***лежат*** */ на гвоздике* ***висят***
локализации	быть, находиться	+	***У нее у рюкзака*** *все карманы были* ***снаружи***
способности	владеть шпагой	+	***У нее у подруги*** *сын пятью языками* ***владеет***
существования	бывать, водиться	+	***У меня у кошки*** *блохи* ***завелись***
отношения	равняться, включать	+	***У тебя у мамы*** *дважды два всегда* ***равняется*** *восемнадцати*
интерпретации	ошибаться	+	***У меня у знакомой*** *сын номером* ***ошибся*** *– вместо аэропорта в поликлинику позвонил*

Таблица 1. Конструкция у Х-а у Y-а и фундаментальная классификация предикатов (по [Апресян 2006])[7]

Актуальное значение	**Возможна ли конструкция**	**Пример**
НСВ		
Актуально-длительное	+	***У меня у девушки*** *мать небось сейчас* ***места себе не находит*** ***У меня у бабушки*** *сейчас небось чайник* ***кипит, свистит*** *во всю мощь, а она и не слышит: над книжкой дремлет.* ***У меня у знакомой*** *сын сейчас выпускное сочинение* ***пишет****, медалист (пример К. Ю. Дружкина).* *Когда я пришел,* ***у меня у сына*** *все книжки* ***лежали*** *под столом, а рюкзак висел на люстре.*
Дуративное	+	***У меня у комп****а клава уже неделю* ***глючит****.*

[7] Примерно те же результаты дает тестирование классификации предикатов, используемой нашей рабочей группой (подробнее см. [Кибрик и др. 2006]).

048-Leontjevy:_Layout 1 18.05.2008 15:45 Page 314

Актуальное значение	Возможна ли конструкция	Пример
НСВ		
Узуальное	+	***У меня у друга*** *родители* ***отдыхают*** *только в Гаграх.*
Потенциальное	+	***У меня у подруги*** *сын уже* ***ходит****.*
Многократное	+	***У меня у соседей*** *кошка по утрам так* ***орет****…* ***У меня у соседки*** *сын каждый день с разбитым носом домой* ***приходит****, а ей хоть бы хны..*
Неактуально-статальное	+	***У меня у друга*** *кот вареные яйца* ***любит****.*
Континуальное	+	***У меня у друга*** *отец на физтехе* ***преподает****…*
Общефактическое Результативное	+	***У меня у друга*** *сын однажды самого Якубовича на телевидение* ***подвозил****.* ***У тебя у друга*** *цыгане деньги* ***воровали****?*
Общефактическе двунаправленное	+	***У нее у сына*** *жена на полгода* ***уезжала****, так она с ним жила – а то бы с голоду помер.*
- нерезультативное	+	***У меня у знакомых*** *девочка туда* ***поступала*** *– провалилась.*
- Непредельное	+	***У меня у подруги*** *сын в детстве мышей* ***боялся*** *– а сейчас разводит их.*
СВ		
Конкретно-фактическое	+	***У нас у соседей*** *кошка* ***пропала****.*
Наглядно-примерное (узуальное)	+	***У нас у соседей*** *кошка, как увидит собаку, сразу на дерево* ***забирается.***
потенциальное	+	***У нас у соседей*** *собака любую дичь* ***догонит****.*
Суммарное	+	***У меня у соседа*** *сын два раза подряд нажал и быстро – сразу открылось все. (пример А. Леонтьева)*

Таблица 2. Конструкция у Х-а у Y-а и аспектуальная классификация предикатов (по [Зализняк, Шмелев 2000])

2.3. Ограничения на участников ситуации

2.3.1. Ограничения на первый посессор. В позиции первого посессора употребляются главным образом личные местоимения (72 из 73 примеров Национального корпуса) – см. Таблицу 3.

Местоимение	Частота встречаемости в Национальном корпусе (%)	Частота встречаемости в корпусе ЭД (%)
1SG	61,6	65,5
2 SG	2,7	9,8
3 SG	8,2	12,7
1 PL	20,5	10,1
2PL	3	0,1
3PL	4	1,8

Таблица 3. Употребление личных местоимений в позиции первого посессора УУК

Как видно из Таблицы 3, наиболее частотна конструкция для местоимений 1-ого лица единственного числа Далее примерно с одинаковой частотой местоимения 1-ого лица множественного числа и 3-его лица единственного числа. Местоимения 2-ого лица единственного и «вежливого» множественного числа сравнительно редки; местоимения 2-ого и 3-его лица множественного числа практически не встречаются.

По понятным причинам в подавляющем большинстве случаев референт первого посессора одушевленный[8]**.** Неодушевленный первый посессор встретился всего в одном примере (22).

(22) (Обсуждается, чем один велосипед лучше/хуже другого) …*Но* ***у него*** (у велосипеда – А.Л.) ***у неведущего колеса*** *поперечная устойчивость хуже, чем у ведущего* …[ЭД].

[8] Ситуации с неодушевленным говорящим (а местоимения 1-ого лица составляют примерно 80% случаев употребления УУК) вещь если не абсолютно невозможная, то по крайней мере достаточно необычная и редкая. В доступных нам корпусах такого материала не встретилось вовсе.

048-Leontjevy:_Layout 1 18.05.2008 15:45 Page 315

Как было сказано, существительные в позиции первого посессора употребляются крайне редко. Единственный пример, встретившийся в Национальном корпусе, содержит существительное в родовом статусе (23).

(23) *...Наши слушатели говорят, что* ***у олигархов у каждого ребёнка*** *на каждый предмеет по своему преподавателю и ещё директор* [ruscorpora].

Конкретно-референтные существительные также возможны, но, как кажется на первый взгляд, подобные примеры выглядят несколько коряво (ср. (24) vs. (25)):

(24) #***У Васи у друга*** *компьютер сломался – он поехал починить.*

(25) - *А что Вася не пришел? -* ***У него у друга*** *компьютер сломался – он чинить поехал.*

Подобного рода корявость, однако, исчезает, как только вместо экспериментального «Васи», до которого реальному слушающему, равно как и говорящему, в сущности, нет никакого дела, появляется имя лица, так или иначе входящего в личную сферу собеседников (ср. реальный диалог (26)).

(26) - *А как «дисплазия» пишется?- Кажется, через «а», а что?- Да* ***у М.Ц. у пса*** *дисплазию обнаружили, а я сомневалась, как в комментах написать…*

Имя М.Ц. упоминается в диалоге впервые, т. е. ранее не было актуализовано, однако его референт хорошо знаком обоим говорящим.

2.3.2. Ограничения на второй посессор. Возможные генитивные отношения. Для УУК существует жесткое ограничение на второй посессор и, соответственно, на смысловое отношение между первым и вторым посессором – генитивное отношение. В конструкции употребляются только пять генитивных отношений:

• отношение родства и/или социальной близости (*у меня у брата/у друга*) – абсолютное большинство примеров;

• собственно посессивное отношение, включая отношение «надетая одежда»[9] (*у меня у кота/ у сумки/у платья…*) – в большой части примеров этой группы позицию второго посессора занимают названия домашних животных, прагматически близкие к именам родства и социальной близости;

• отношение части-целого, включая части тела (*у него у неведущего колеса / у меня у этого зуба кариес* [ЭД]*; у меня у пальца на ноге косточка болит* [ЭД]);

• отношение «автор – продукт» (*У меня у статьи название не придумывается*).

Другие генитивные отношения, даже те, в которых посессор одушевленный, как кажется, невозможны: **у нас у бригады…; *у него у бригады…; *у них у вигвама[10], *у него у жертвы…; *у тебя у появления…; *у них у избиения…; * у нее у красоты…; *у вас у услуг…*

2.4. Ограничения, связанные со структурой дискурса. Как уже говорилось, конструкция употребляется в устной разговорной речи или ее аналоге – электронном дискурсе.

2.4.1. Диалогическая речь. Данный тип дискурса для рассматриваемой конструкции предпочтителен.

В иллокутивно вынуждающих репликах УУК употребляется в сравнительно небольшом количестве коммуникативных ситуаций. Это вопрос-просьба о помощи или сообщение новостей ((27), (28)). Здесь конструкция служит для описании ситуации, на которую собеседник должен отреагировать, и всегда кодирует новую тему.

(27) *Добрый день! Подскажите, пожалуйста, как правильнее поступить в следующей ситуации:* ***у меня у бабушки приватизированная квартира*** *(приватизировали в прошлом году, собственник-бабушка). В этой квартире прописаны бабушка, соответственно, и еще я. Как переоформить квартиру на меня (наследство, дарение, купля-продажа?), чтобы впоследствии квартира перешла в мою собственность и это не могло быть оспорено?* [ЭД].

(28) *Девушки!!! Ура!!!!* ***У меня у доченьки*** *первый зубик вылез!!* [ЭД].

В иллокутивно вынужденных репликах выделяются две основные коммуникативные ситуации. Во-первых, это ответ на вопрос или претензию. Обычно такая реплика имеет сложную структуру и состоит более чем из одного предложения. Высказывание с УУК вводит объяснение причины сказанного, доказательство или пример, подтверждающий правоту говорящего ((29) – (32)). Во-вторых, – ответ на вопрос или просьбу о помощи, содержащий описание опыта отвечающего, подобного ситуации, описанной спрашивающим (32). Для такого ответа практически обязательны маркеры типа *так же, такой же, тоже*.

(29) [Интервьюер]. *Ну а тогда / в 30-е годы / или / было время / оставалось на чтение?* [Респондент1]. *Было.* ***У нас у соседей*** *было очень хорошее издание Толстого.* [Ruscorpora].

(30) *— У вас есть место, где вы можете отсидеться несколько дней так, чтобы Цуладзе не смог вас найти? — Наверное, я найду.* ***У меня у тети*** *дача в Малаховке*. [Ruscorpora].

(31) (Ответ на претензию приятеля, не получившего обещанное письмо по электронной почте) *Извенясь***,**

[9] Подробнее об отношении «надетая одежда» см. [Леонтьев 2005], [Кибрик и др. 2006].

[10] Конструкция возможна только как реализация посессивного отношения, но не отношения «обитатель – место обитания».

048-Leontjevy:_Layout 1 18.05.2008 15:45 Page 316

у меня у прова *мыльник загнулся, ежли не получил перепослал* [ЭД].

(32) (Ответ на недоумение собеседника, возникшее оттого, что пользователь вдруг стал использовать не свое имя – «ник»). *Maple, это* ***у меня у брата*** *ник SIPO, я забыл выйти из его пользователя* [ЭД].

(33) (Обсуждается некоторая школа, ее достоинства и недостатки). ***У меня у друга сын*** *в этом году в первый класс* ***туда*** *ходит. Чтобы поступить туда желательно годик походить на подготовительное отделение :). Они в целом довольны* [ЭД].

2.4.2. Монологическая речь. Другой тип дискурса, привычный для УУК, – нарратив. УУК обычна для его начала – введения референта, новой темы ((34), (35)).

(34) *Ещё одна игрушка / про которую я вам хотел рассказать / это была такая история / довольно печальная / но бытовая история.* ***У меня у приятеля сошла с ума жена*** */ и мне пришлось за ней какое-то время поухаживать / но всё равно она потом погибла. И когда это произошло / а для меня это было очень серьёзно.* [Ruscorpora].

(35) (Обсуждается тема «хотите, верьте, – хотите, нет»; рассказываются разные истории). ***А у меня у знакомого*** *крестный купил авто... и ящик водки - поехали домой обмывать. Водку оставили до вечера в багажнике, потом пошли за ней, а машины и водки нет... через несколько дней нашли эту машину а в ней пару трупов... угнали и той водкой и потравились... вот и не верь в судьбу...* [ЭД].

3. Предназначение конструкции с рекурсивным *у*. Попытка объяснения ограничений. Как известно, русские конструкции с одинарным *у*, по крайней мере употребляемые в начале предложения, кодируют топик ((36), (37))[11].

(36) ***У слонов*** *уши большие.*

(37) *А* ***у Васьки Иванова*** */ собака вчера милиционера покусала.*

Топик, как правило, соотносится с актуализованной (данной) или с легко актуализуемой (доступной) информацией.

В конструкции с рекурсивным *у* второй посессор функционально тот же топик, что и в одинарных *у*-конструкциях. Но, в отличие от таких «классических» топиков, он для слушателя, как правило, абсолютно нов и никак не может быть актуализован. Чтобы актуализовать его, и вводится второе *у,* которое вводит референт, доступный слушающему, с одной стороны, и связанный с референтом нового топика, с другой.

Данное свойство конструкции благоприятствует ее употреблению в перечисленных выше коммуникативных ситуациях и объясняет существующее ограничение на первый актант.

В самом деле, конструкция, способная вводить любой референт, даже ранее не актуализованный, прекрасно подходит для начала нарратива или описания ситуации. При объяснении же причины и описании опыта посессор, структурно необходимый говорящему, оказывается для слушающего абсолютно «чужим» и, следовательно, нуждающимся в актуализации.

Местоимения, особенно первого лица, всегда кодируют актуализованную информацию. То же можно сказать об именах лиц, находящихся в личной сфере собеседников.

Ограничение на второй посессор и зависимость расщепления конструкции семантики связи между ее компонентами можно было бы объяснить разной степенью слитности (или, в других терминах, когнитивной сопряженности) генитивных групп. Так, отношения родства и социальной близости, прототипические для рассматриваемой конструкции, оказываются максимально слитными (сопряженными). Иными словами, имена родства и социальной принадлежности обладают наиболее сильной валентностью на посессор. Менее слитным оказывается собственно посессивное отношение. Еще менее – отношения «часть – целое» и «автор – продукт».

По всей вероятности, иерархия слитности генитивных отношений, выявленная на материале конструкции с рекурсивным *у*, проявит себя и в других областях языка.

Список литературы

1. Апресян Ю.Д. Фундаментальная классификация предикатов // Апресян В.Ю., Апресян Ю.Д., Бабаева Е.Э., Богуславская О.Ю., Иомдин Б.Л., Крылова Т.В., Левонтина И.Б., Санников А.В., Урысон Е.В. Языковая картина мира и системная лексикография. – М., 2006.

2. Брыкина М.М. Идентификация посессивных связей для имен, обозначающих части тела, в русском языке // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'2007». – М., 2007.

3. Зализняк Анна А., Шмелев А.Д. Введение в русскую аспектологию. – М., 2000.

4. Кибрик А.Е. Внешний посессор как результат расщепления валентностей // Слово в тексте и словаре: Сб. ст. к семидесятилетию академика Ю.Д. Апресяна. – М., 2000.

[11] Подробнее см. [Мельчук 1995], [Падучева 2004], [Кибрик 2000] с дальнейшей библиографией.

048-Leontjevy:_Layout 1 18.05.2008 15:45 Page 317

5. Кибрик А.Е. Когнитивная сопряженность и организация языковой структуры // Вторая международная конференция по когнитивной науке. Тезисы докладов. СПб, 2006. Том 1.

6. Кибрик А.Е., Брыкина М.М., Леонтьев А.П., Хитров А.Н. 2006. Русские посессивные конструкции в свете корпусно-статистического исследования // Вопросы языкознания, М.: Наука.2006.

7. Леонтьев А.П. Влияние типа генитивного отношения на конструкции с внешним посессором в русском языке // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'2005». – М., 2005

8. Леонтьев А.П., Леонтьева А.Л. Еще раз к вопросу о семантике генитивных отношений // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2006» (Бекасово, 31 мая – 4 июня 2006 г.). М., 2006.

9. Леонтьев А.П. Семантические и синтаксические свойства именной группы в перспективе конструкций с внешним посессором. – автореф. дис. канд. филол. наук. – М., 2008.

10. Мельчук И.А. 'Glaza Maši golubye vs. Glaza u Maši golubye: Choosing between Two Russian Constructions in the Domain of Body Parts // Мельчук И.А. Русский язык в модели «Смысл ó Текст». – Москва – Вена, 1995.

11. Падучева Е.В. Расщепление генитивной группы; диатезы с внешним Посессором // Падучева Е.В. Динамические модели в семантике и лексике.

12. Хитров А.Н. Посессивная конструкция и семантика неучастия в ситуации // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'2005». – М.,2005.

13. Циммерлинг А.В. Обладать и быть рядом // Логический анализ языка. Языки пространств. М., Языки русской культуры, 2000.

049-Litvinenko_AO:_Layout 1 18.05.2008 15:45 Page 318

КОНСТРУКЦИИ С ДВОЕТОЧИЕМ В УСТНОМ НАРРАТИВЕ: ПРОБЛЕМЫ ТРАНСКРИБИРОВАНИЯ[1]

ASYNDETON AND COLON. TRANSCRIBING SPOKEN NARRATIVE

Литвиненко А.О. *(allal1978@rambler.ru)*
Московский государственный университет им. М.В. Ломоносова

Работа посвящена проблемам транскрибирования закрытого класса бессоюзных конструкций в устном нарративе, которым в письменном языке соответствуют бессоюзные сложные предложения с двоеточием.

1. Введение

В исследованиях русского синтаксиса, с одной стороны, и русской устной речи, с другой, неоднократно отмечалось, что бессоюзное сложное предложение является ярко выраженным признаком разговорной речи (см., например, Пешковский 1928; Лаптева 1976; Шведова 1980; Белошапкова 1989; Граудина, Ширяев 1999; Валгина 2000 и др.). Между тем, несмотря на обширнейшую историю вопроса и длительный спор о природе бессоюзия и возможности или невозможности соотнесения бессоюзных конструкций с союзными, многие весьма распространенные конструкции этого рода еще изучены недостаточно. Интонационные средства выражения бессоюзной связи редко рассматриваются подробно, поскольку либо априорно постулируются как основные и в дальнейшем практически не рассматриваются (см. работы А.М. Пешковского, А.Н. Гвоздева, О.А. Лаптевой, Н.С. Валгиной и др.), либо отвергаются в качестве таковых (позиция Н.С. Поспелова, Е.Н. Ширяева, В.А. Белошапковой, с одной стороны, и С.О. Карцевского, И.Н. Кручининой – с другой). Пунктуация же как средство структурирования текста, а также выражения смысловых связей между частями сложного предложения и интонации и вовсе зачастую остается полностью за пределами исследований.

Соотношение пунктуации и интонации не является однозначным: часть пунктуационных знаков ставится исключительно по грамматическим правилам, часть – по смысловым, часть – по интонационным; чаще всего наблюдается триединство грамматики, семантики и интонации (Валгина 2004). Некоторые типы бессоюзных конструкций совершенно разной семантики произносятся приблизительно с одной и той же интонацией (или, по крайней мере, различить их весьма непросто); тем более очевидно, что одну и ту же написанную фразу можно произнести с разными интонациями. Столь же очевидно, что некоторые конструкции допускают вариативность и при произнесении, и при записи, практически не меняя при этом смысла. Однако существуют пунктуационные и интонационные средства, не только обладающие вполне конкретной семантикой, но и соотносящиеся друг с другом довольно жестко, что позволяет рассматривать их как своеобразные языковые знаки с варьирующимся в зависимости от канала связи означающим.

В данной работе к рассмотрению предлагаются конструкции, использующие в качестве средства поддержания связи между частями бессоюзного предложения **двоеточие** на письме и соответствующую ему «**интонацию двоеточия**» (за неимением лучшего термина) в устной речи. В качестве материалов используется корпус детских рассказов о сновидениях, собранных кандидатом медицинских наук Е.А. Корабельниковой, предложившей метод экспертного анализа сновидений для диагностики и терапии невротических заболеваний у детей и подростков. Это 129 рассказов детей в возрасте от 7 до 17 лет.

2. Семантика двоеточия

Двоеточие – пунктуационный знак, являющийся, аналогично запятой (а также точке с запятой и тире), выражением определенного вида дискурсивной незавершенности и одновременно связи между частями синтаксического целого. В этом вышеперечисленные знаки противопоставлены целой группе других, являющихся выражением определенного типа иллокуции (точка, вопросительный и восклицательный знаки и многоточие).

В справочнике Розенталь 1967/2004 приводятся следующие случаи постановки двоеточия «между частями бессоюзного сложного предложения, распадающегося на две части» (с. 160):

[1] Данная работа является частью коллективного исследования, посвященного проблемам транскрибирования устного нарративного дискурса, проводимого совместно группой исследователей под руководством А.А. Кибрика и В.И. Подлесской. Работа поддерживается грантом РФФИ 06-06-80470.

049-Litvinenko_AO:_Layout 1 18.05.2008 15:45 Page 319

(i) «если вторая часть (одно или несколько предложений) разъясняет, раскрывает содержание первой части (между обеими частями можно вставить *а именно*)»;
(ii) «если в первой части посредством глаголов *видеть*, *смотреть*, *слышать*, *понимать*, *узнать*, *чувствовать* и т.п. делается предупреждение о том, что далее последует изложение какого-либо факта или какое-нибудь описание (в этих случаях между частями обычно можно вставить союз *что*)»;
(iii) «если в первой части имеются глаголы *выглянуть*, *оглянуться*, *прислушаться* и т.п., а также глаголы со значением действия, предупреждающие о дальнейшем изложении и допускающие вставку после себя слов *и увидел, что*, *и услышал, что*, *и почувствовал, что* и т.п.»;
(iv) «если вторая часть указывает основание, причину того, о чем говорится в первой части (между обеими частями можно вставить союз *потому что*, *так как*, *поскольку*)»;
(v) «если вторая часть представляет собой прямой вопрос» (с. 160-161).

Из приведенных цитат видно, что случаи использования двоеточия сводятся к довольно ограниченному количеству семантических контекстов, при этом подразумевают довольно жесткую двухчастную структуру бессоюзного сложного предложения. Однако очевидным недостатком сформулированных правил является чрезмерная привязанность их к конкретным лексемам и постоянное уподобление рассматриваемых бессоюзных конструкций сложноподчиненным предложениям, чего хотелось бы избежать: как уже упоминалось выше, такое уподобление представляется довольно спорным само по себе.

В специальной работе В.И. Подлесской (1987), где двоеточие рассматривается как ряд омонимичных единиц, приводится следующий список его употреблений.

1) Вторая часть конструкции с двоеточием (после знака) является в некотором смысле кореферентной первой части, дополняя или раскрывая ее содержание: «X, а именно: У». Это значение может принимать различные формы: обобщающее слово + список, общее утверждение + частный случай и т.д.

2) Вторая часть конструкции с двоеточием указывает на причину или обоснование действия или ситуации, описанной в первой части.

3) Вторая часть конструкции с двоеточием указывает либо на информацию, полученную в результате описанного в первой части наблюдения или действия, сделавшего возможным наблюдение (*вижу: идет; выглянул в окно: солнце светит*), либо на содержание некоего сообщения. Двоеточие часто вводит содержание при предикатах восприятия, речи, мысли и т.д.

Очевидно, что использование двоеточия при прямой речи, в том числе упомянутый у Д.Э. Розенталя «прямой вопрос», является частным случаем ввода информации, описанного у В.И. Подлесской в пункте 3. Дополнительными выразителями связи этого типа могут являться определенные лексические маркеры: *вот что, вот такой, так, в частности* и т.п. В дальнейшем мы будем брать за основу именно этот список значений, как более обобщенный и менее привязанный к конкретным лексемам.

Что же объединяет эти три разных смысла? Почему для них выбирается общая форма выражения – маркированная форма незавершенности, в противоположность немаркированной форме – запятой? По всей вероятности, следует признать, что объединяющим элементом служит наличие у первой части незаполненной предикативной семантической валентности. Используя данную конструкцию, говорящий (пишущий) указывает не просто на незавершенность первой части конструкции, но на ее семантическую неполноту, которую обещает заполнить во второй части.

3. «Интонация двоеточия» как фиксированная интонационная конструкция

Еще А.М. Пешковский отмечал «объяснительную» и «предупредительную» интонацию как четко выраженные, фиксированные типы, приводя в пример конструкции, подпадающие под описания значений 2) и 1) по В.И. Подлесской соответственно. О тех же типах интонации как о хорошо выраженных пишут А.Н. Гвоздев (1949), Н.С. Валгина (2000, 2004). А.Н. Гвоздев также подчеркивает сходство «предупредительной» («изъяснительной») интонации (в его терминах – мелодики) и интонации пояснительной, а также сходство с ними интонации ввода прямой речи и обобщения перед списком частных случаев.

Явная обособленность и индивидуальность этой интонации заставили нас задуматься о необходимости ее системного отображения в разрабатываемой транскрипционной системе. Перцептивное ее вычленение в большинстве случаев не представляет труда – именно поэтому ее чаще всего приводят в качестве примера фиксированной интонационной конструкции русского языка. Однако такое решение приводит к закономерному вопросу: что именно представляет собой «интонация двоеточия» с точки зрения просодии? А.М. Пешковский (1918) характеризует ее как связанную с понижением тона, резким ударением в первой части и обязательностью паузы между частями. Действительно, при лабораторном прочтении фраз с двоеточием различной семантики эти признаки, особенно непременная пауза, очень хорошо заметны. Однако при исследовании примеров из корпуса живой речи, где соответствующие фразы вписаны в контекст нарратива, транскрайбер сталкивается с различными трудностями.

В нашем корпусе 64 случаев использования «интонации двоеточия» приходится на ввод чужой речи и всего 23 – на остальные его значения. Прежде всего, оказывается, что в некоторых случаях тому, что на слух

049-Litvinenko_AO:_Layout 1 18.05.2008 15:45 Page 320

воспринимается как «интонация двоеточия», соответствует ровный тон или сложные разновидности: падение в ровный; «крышка»; ровный тон, переходящий в падение. Семантика, впрочем, не позволяет сомневаться в принадлежности этих конструкций к рассматриваемому классу. Также ни при каких обстоятельствах здесь не используется настоящий подъем (ИК-3).

В подавляющем большинстве случаев (54 и 21 соответственно) «интонация двоеточия» действительно представляет собой падение, но его характер не так однороден и четко выражен, как в изолированных примерах. Резкость этого падения варьируется от случая к случаю и от говорящего к говорящему. Кроме того, на вид тональной кривой значительно влияет местоположение акцентоносителя в составляющей с двоеточием (впрочем, это хорошо видно и в сконструированных примерах, специально записанных для проверки данных корпуса). Если акцентоноситель расположен в конце составляющей, а тем более если в нем нет заударных слогов (как в лексемах *сказал*, *говорит*, *говорю, так, там* и т.п.), то рассматриваемая интонация обычно представляет собой типическую ИК-1, с той разницей, что целевая частота падения на 1-3 полутона выше типичной точки для данного говорящего.

(1). *..(0.1) Затем \видовцы /пришл<u>и</u>, ...(0.6) /и-и \перерисов<u>а</u>ли его, ..(0.4) очень не \т<u>а</u>к: ..(0.4) /они \перерисов<u>а</u>ли <<u>его</u>>, у него были щуритые /глаза ..(0.4) /нос ...(0.6) \неулыбающийся /рот ....(1.0) /брови ...(0.6) и вместо ляг= ..(0.3) /лягушки ..(0.2) \ш<u>и</u>шка.*

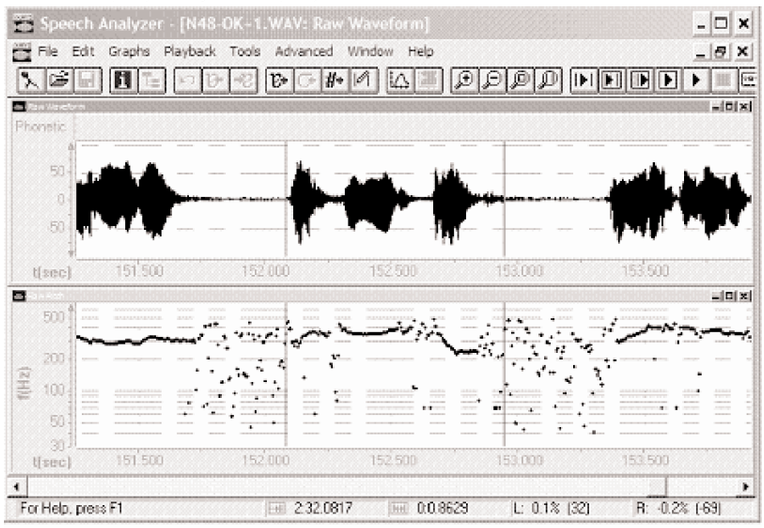


Рис. 1. Тонограмма фрагмента примера (1). Курсорами выделена составляющая «очень не \так».

При сдвижении акцентоносителя влево, в середину составляющей, у тональной кривой нередко появляется «хвост» – заударный подъем.

(2). *Но /**я** увидел в-во /\сне ..(0.1) совсем \друг<u>о</u>го, ...(0.5) не \так<u>о</u>го ↑как они: /\уродство одно и /\т<u>о</u> же, но-о ..(0.4) /\в<u>о-о</u>лосы другие_*

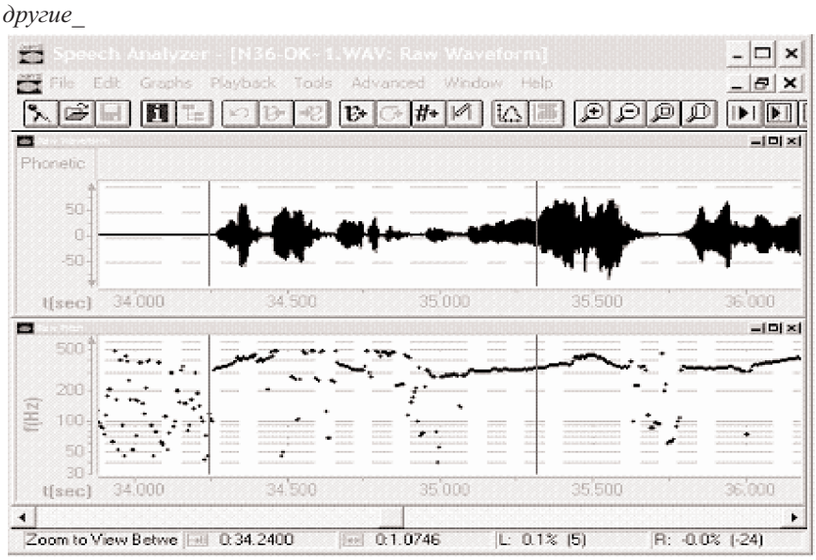


Рис. 2. Тонограмма фрагмента примера (2). Курсорами выделена составляющая «не \такого ↑как они».

049-Litvinenko_AO:_Layout 1 18.05.2008 15:45 Page 321

При дальнейшем сдвижении акцентоносителя к началу составляющей у тональной кривой нередко появляется второе падение, дублирующее первое, но более слабое.
(3). *эээ(0.5) а где-то в /середине ...(0.9) {ЦОКАНЬЕ 0.3} ..(0.2) такой как бы \кадр встаёт: ...(0.5) что-о ..(0.4) как бы ..(0.3) /\я /лежу-у,*

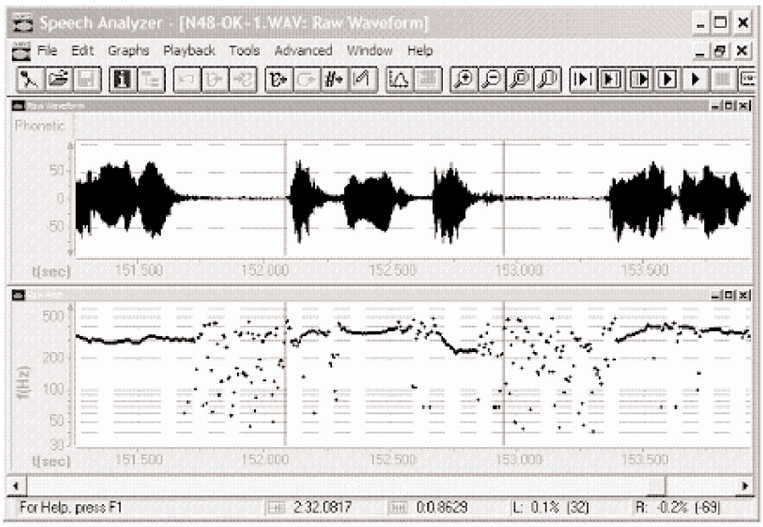


Рис. 3. Тонограмма фрагмента примера (3). Курсорами выделена составляющая «\кадр встаёт».

Что касается паузы на границе между частями бессоюзной конструкции, соединяемой двоеточием, то и она частотна, но отнюдь не обязательна. Так, в примерах (1) и (3) пауза присутствует, а в примере (2) она не только отсутствует, но и вся составляющая произносится слитно с последующим контекстом. Та же ситуация нередко встречается и при введении прямой речи.

Таким образом, ни один из перечисленных просодических признаков (за исключением отсутствия значимого подъема в главном акценте составляющей), строго говоря, не обязателен для создания «интонации двоеточия», но в сумме они создают впечатление «правильно» произносимой конструкции. В сочетании с наличием семантической связи из описанного выше закрытого списка, такая интонация заставляет слушающего ожидать обязательного продолжения и заполнения предикативной валентности, на которую указывает двоеточие. Используя «интонацию двоеточия», говорящий сужает круг возможных интерпретаций связи между уже сказанным и ожидаемым, еще не произнесенным контекстом, таким образом облегчая восприятие текста слушающим.

4. Заключение

По материалам нашего корпуса можно сделать следующий вывод: «конструкции с двоеточием» представляют собой сложный интонационно-семантический комплекс. Просодические признаки таких конструкций гибки и варьируются в пределах определенного диапазона; семантика также описывается в виде списка признаков и, как кажется, не сводима к одному значению. Однако сочетание семантики и интонации практически всегда позволяет адресату однозначно идентифицировать связь между частями конструкции как относящуюся к описываемому классу – и для выражения этой связи в письменном языке и служит двоеточие. Стабильность и четкая очерченность этого класса конструкций позволяет ввести их в транскрипционную систему русской разговорной речи.

Список литературы

1. Белошапкова В.А. (ред.). Современный русский язык. М.: Высшая школа, 1989.
2. Валгина Н.С. Актуальные проблемы современной русской пунктуации. М.: Высшая школа, 2004.
3. Валгина Н.С. Синтаксис современного русского языка: учебник. М.: Агар, 2000.
4. Валгина Н.С., Розенталь Д.Э., Фомина М.И. Современный русский язык. М.: Логос, 2006.
5. Гвоздев А.Н. О фонологических средствах русского языка. М.: АН СССР, 1949.
6. Гвоздев А.Н. Очерки по стилистике русского языка. М.: Просвещение, 1965.

049-Litvinenko_AO:_Layout 1 18.05.2008 15:45 Page 322

7. Граудина Л.К., Ширяев Е.Н. (ред.). Культура русской речи. Учебник для вузов. М.: ИРЯ им. В.В. Виноградова, 1999.

8. Карцевский С.О. Бессоюзие и подчинение // «Вопросы языкознания», №2, М.: 1961.

9. Лаптева О.А. Русский разговорный синтаксис. М.: 1976.

10. Пешковский А.М. Русский синтаксис в научном освещении. М.: 1928.

11. Пешковский А.М. Школьная и научная грамматика. М.: 1918.

12. Подлесская В.И. Опыт семантического анализа знака препинания (на примере двоеточия в русском языке) // Научно-техническая информация. Серия 2, 12. М.: 1987. С. 25-28.

13. Поспелов Н.С. О грамматической природе и принципах классификации бессоюзных сложных предложений //Вопросы синтаксиса современного русского языка. М.: 1950. С. 25-37.

14. Розенталь Д.Э. Справочник по правописанию и литературной правке. 9-е изд. (под ред. И.Б. Голуб). М.: Айрис-пресс, 2004.

15. Шведова Н.Ю. (ред.). Русская грамматика, т. I, II. М.: Наука, 1980.

050-Lobanov:_Layout 1 18.05.2008 15:45 Page 323

АЛГОРИТМ СЕГМЕНТАЦИИ ТЕКСТА НА СИНТАКСИЧЕСКИЕ СИНТАГМЫ ДЛЯ СИНТЕЗА РЕЧИ

AN ALGORITHM OF TEXT SEGMENTATION ON SYNTACTIC SYNTAGMAS FOR TTS SYNTHESIS

Лобанов Б.М. *(lobanov@newman.bas-net.by)*
Объединенный институт проблем информатики НАН Беларуси, Минск, Беларусь

Предлагается алгоритм сегментации текста на синтаксические синтагмы, основанный на анализе устойчивых фразеологических и грамматико-смысловых словосочетаний, составляющих предложение. Основной смысл выделения в предложении рассматриваемых словосочетаний заключается в том, что теперь свобода его разделения на синтагмы ограничивается, а именно: граница синтагмы может находиться только за пределами словосочетаний, но не внутри их.

Введение

Просодическая разметка текста заключается в его членении на синтагмы, разметке синтагм на акцентные единицы и маркировке интонационного типа синтагм [1].

Под синтагмой понимается самостоятельная в интонационном смысле часть предложения или всё предложение. Установка границ синтагм влияет на передачу интонационных характеристик при синтезе речи, а также на передачу смыслового содержания. При разбиении текста на синтагмы важно не поставить границу синтагмы там, где она может нарушить смысловое восприятие речи (или передачу смыслового содержания текста), например, между предметом и его признаком. Для установки границ синтагм при синтезе речи по тексту используются определённые правила синтагматического членения, базирующиеся на синтаксическом анализе текста, на учёте фактора речевого дыхания [2], а также на статистическом анализе особенностей синтагматического членения естественной речи конкретного диктора [3LobTs].

Первым этапом просодической разметки текста является его членение на пунктуационно-лексические синтагмы [4Present]. Пунктуационно-лексической синтагмой (ПЛС) считается всё предложение (при отсутствии в нём знаков препинания) или часть предложения, ограниченная любым знаком препинания или каким-либо из лексических маркеров. Следует заметить, что даже после разбиения предложения на ПЛС их длина может оказаться всё же слишком большой.

Пример:

«*Но молодая жена упорно продолжала отстирывать белую в кровавых пятнах рубаху мужа посиневшими от холода руками в железном тазике с ледяной водой*».

Очевидно, что при отсутствии механизма дальнейшего членения таких предложений на более мелкие синтаксические синтагмы (СС) неизбежно возникнут затруднения в понимании смысла синтезированной речи. Идеальным решением проблемы дальнейшего членения такого рода ПЛС на СС было бы использование комплекса правил их глубинного синтаксического разбора [5Bogus]. Однако, в виду сложности и недостаточной разработанности таких правил, в данной работе предлагается использование процедуры поверхностного синтаксического анализа, опирающейся на доступную морфосинтаксическую информацию о словосочетаниях, составляющих ПЛС.

1. Общая структура алгоритма сегментации на синтаксические синтагмы

Словосочетание рассматривается в грамматике как пара по смыслу и грамматически связанных слов, выделяемая из предложения [6Gramm]. Являясь наряду со словом элементом построения предложения, словосочетание выступает в качестве одной из основных синтаксических единиц. Непосредственной целью

050-Lobanov:_Layout 1 18.05.2008 15:45 Page 324

рассматриваемой процедуры поверхностного синтаксического анализа является предварительное разбиение ПЛС на последовательность словосочетаний 2-х типов: устойчивые фразеологические словосочетания (ФЛС) и грамматико-смысловые словосочетания (ГСС). Основной смысл выделения в ПЛС словосочетаний типа ФЛС и ГСС заключается в том, что теперь свобода разделения ПЛС на СС ограничивается. Граница синтагмы может находиться только за пределами ФЛС или ГСС, но не внутри их.

Предлагаемая общая структура процедуры просодической разметки на синтаксические синтагмы для синтеза речи по тексту, основанная на анализе словосочетаний, представлена на рис. 1.

Рис. 1. Общая структура алгоритма сегментации на синтаксические синтагмы

Рассмотрим подробно функционирование каждого из блоков алгоритма, представленного на рис.1.

2. Выделение фразеологических словосочетаний

В анализируемой синтагме отмечаются ФЛС, найденные в словаре устойчивых словосочетаний. К фразеологическим словосочетаниям относятся [6]:

– фразеологические сращения – «*попасть впросак», «бить баклуши», «ничтоже сумняшеся», «собаку съесть»* и др.

– фразеологические единства – «*зайти в тупик», «бить ключом», «плыть по течению», «брать в свои руки», «прикусить язык»* и др.

– фразеологические сочетания – «*потупить взор»,* «*щекотливый вопрос»,* «*бархатный сезон»*, «*поголовные аресты»* и др.

Выделяются следующие типы компонентного состава фразеологизмов:

– сочетание прилагательного с существительным: *краеугольный камень, заколдованный круг, лебединая песня*;

– сочетание существительного в именительном падеже с существительным в родительном падеже: *точка*

050-Lobanov:_Layout 1 18.05.2008 15:45 Page 325

зрения, камень преткновения, бразды правления, яблоко раздора;

– сочетание имени существительного в именительном падеже с существительными в косвенных падежах с предлогом: *кровь с молоком, душа в душу, дело в шляпе*;

– сочетание предложно-падежной формы существительного с прилагательным: *на живую нитку, по старой памяти, на короткой ноге*;

– сочетание глагола с существительным (с предлогом и без предлога): *окинуть взором, посеять сомнения, взять в руки, взяться за ум, водить за нос*;

– сочетание глагола с наречием: *попасть впросак, ходить босяком, видеть насквозь*;

– сочетание деепричастия с именем существительным: *спустя рукава, скрепя сердце, сломя голову*.

Позиции слабых и сильных словесных ударений в устойчивых сочетаниях могут быть определены в словаре сочетаний, при этом одно из слов обязательно несёт сильное ударение. При отсутствии помет слабых и сильных ударений вполне допустима установка сильного ударения на каждом из слов устойчивого словосочетания.

3. Объединение слов в грамматико-смысловые словосочетания

В зависимости от того, какое слово является первым в словосочетании, различаются основные лексико-грамматические группы словосочетаний. Классификация словосочетаний по признаку первого слова может быть представлена следующей схемой.

Группа прилагательных словосочетаний. Эта группа включает прилагательные, местоимения-прилагательные, порядковые числительные и причастия, которые сочетаются:

• С существительным (*полезная книга, зелёную листву, свою находку, унесённые ветром)*.

Признаки словосочетания: прилагательное + существительное (в одном падеже).

• С инфинитивом (*способный работать, готовый учиться)*.

Признаки: прилагательное + инфинитив.

Группа наречных словосочетаний. Эта группа сочетается:

• С инфинитивом (*безнаказанно игнорировать, хорошо петь).*

Признаки: наречие + инфинитив.

• С наречием (*очень удачно, по-прежнему хорошо).*

Признаки: наречие + наречие

• С существительным *(далеко от дома, наедине с сыном, незадолго до экзаменов).*

Признаки: наречие + существительное (в косвенном падеже с предлогом).

• С местоимением-существительным (*недалеко от них, наедине с ней, незадолго до неё)*.

Признаки: наречие + местоимение-существительное (в косвенном падеже с предлогом).

Группа глагольных словосочетаний. Глагольная группа сочетается:

• С инфинитивом (*предложил выучить, просит взять)*.

Признаки: глагол (в любой форме) + глагол (инфинитив).

• С деепричастием (*идёт оглядываясь, говорить улыбаясь)*.

Признаки: глагол (в любой форме) + деепричастие.

• С наречием (*поступал справедливо, заниматься вдвоем)*.

Признаки: глагол (в любой форме) + наречие.

• С существительным (*искать покоя, писал брату, стоять у дороги, подъехал к дому, встретиться с друзьями)*.

Признаки: глагол (в любой форме) + существительное в косвенном падеже.

• С местоимением-существительным **(***искать их, писал ему, стоять около неё, подъехал к нему, встретился с ними)*.

Признаки: глагол (в любой форме) + местоимение-существительное в косвенном падеже.

Группа числительных словосочетаний. Эта группа сочетается:

• С существительным (*две книги, оба друга, трое в шинелях, сто рублей)*.

Признаки: количественное или собирательное числительное + существительное, в одном падеже.

Группа существительных словосочетаний. Эта группа включает существительные и местоимения-существительные и сочетается:

• С существительным (*письмо родителям, его доклада, оценку выступления, входом в театр)*.

Признаки: существительное + существительное (в отличающихся падежах без предлога или с предлогом).

• С наречием (*прогулка верхом, судак по-польски)*.

Признаки: существительное + наречие.

050-Lobanov:_Layout 1 18.05.2008 15:45 Page 326

Перечисленные правила объединения слов в ГСС представлены в таблице 1, где по горизонтали расположены типы групп словосочетаний в порядке степени (силы) его связности со вторыми в паре словами. В таблице указано также место предпочтительной установки полного (+) и частичного (=) ударений.

Тип словосочетания / Второе слово в паре		Прилагательное	Наречное	Глаголь-ное	Числи-тельное	Существи-тельнное
		1	**2**	**3**	**4**	**5**
Инфинитив	**1**	(=) (+)	(=) (+)	(=) (+)	–	–
Деепричастие	**2**	–	–	(=) (+)	–	–
Наречие	**3**	–	(+) (=)	(=) (+)	–	(=) (+)
Существительн	**4**	(=) (+)	(+) (=)	(+) (=)	(+) (=)	(+) (=)
Местоимение	**5**	–	(+) (=)	(+) (=)	–	–

Таблица 1. Правила объединения слов в словосочетания

Из таблицы видно, что в соответствии с правилами русской грамматики [3], допустимыми и наиболее частотными являются 14 различных типов ГСС. С учётом этого предлагается следующая последовательность действий по разметке синтагм на словосочетания.

1. В синтагме отыскиваются пары слов - прилагательные словосочетания, состоящие из слова группы прилагательных и стоящего справа от него существительного либо инфинитива глагола. Эти пары слов объединяются в словосочетания. Если такой пары не находится, то слово из группы прилагательных остаётся «одиноким».

2. Затем в синтагме рассматриваются оставшиеся слова, т.е. не объединённые в словосочетания по п. 1, и отыскиваются пары слов - наречные словосочетания, состоящие из двух наречий или наречия и стоящего справа от него инфинитива глагола, либо существительного или местоимения-существительного с предлогом. Если таковые находятся, то они объединяются в словосочетания, если нет, то наречие остаётся «одиноким».

3. Далее в синтагме рассматриваются оставшиеся слова, и отыскиваются пары слов - глагольные словосочетания, т.е. глагол в любой форме и стоящие справа от него наречие, инфинитив или деепричастие, которые объединяются в одно словосочетание. Если таковых не обнаружено, то глагол может быть объединён с существительным или с местоимением-существительным в косвенном падеже, стоящим справа, Если их нет, то глагол остаётся «одиноким».

4. В оставшихся необъединённых словах ищутся пары слов - числительные словосочетания, состоящие из количественного или собирательного числительного и стоящего справа от него существительного, согласованного с числительным по падежу, которые объединяются в словосочетания. Если нет, то числительное остаётся «одиноким».

5. Наконец, в оставшихся необъединённых словах ищутся соседние пары слов - существительные словосочетания, состоящие из слова группы существительных и стоящего справа от него наречия либо существительного или местоимения-существительного, которые объединяются в словосочетания. Если таких слов не находится, то существительное остаётся «одиноким».

Ниже в примере (3) показана разметка 4-х различных ПС на словосочетания в соответствии с предложенной последовательностью действий (см. п.п. 1 – 5). Словосочетания отмечены квадратными скобками, в круглых скобках указан тип словосочетания (см. таблицу 1), буквой *Ъ* обозначено объединение знаменательных и служебных слов в фонетическое слово.

(3) Пример:

Если Вам [необходимо активировать(2)] [услугу передачи(5)] даных дляЪвашего [мобильного номера(1)].

Но благодаря [разумному сочетанию(1)] лекарств он [смог остановить(3)] [развитие болезни(5)] [вЪбольшинстве случаев(5)].

Тогда тарификация [Ваших звонков(1)] [начинается сЪмомента(3)] соединения [сЪ телефоном абонента(5)].

[Идеальным решением(1)] [проблемы членения(5)] [такого рода(1)] предложений наЪсинтагмы [былоЪбы использование(3)] [комплекса правил(5)] разбора [наЪсинтаксические компоненты(1)].

050-Lobanov:_Layout 1 18.05.2008 15:45 Page 327

4. Расширение двухсловных ГСС до трёх- и более словных сочетаний

Если ПЛС, обработанная в соответствии с указанной выше в п.п. 1 – 5 последовательностью действий по разметке синтагм на словосочетания, содержит слова, не вошедшие в созданные двухсловные сочетания, то рассматривается возможность расширения до трёх- и более словных сочетаний по следующей схеме.

1. Рассматриваются полученные «прилагательные словосочетания».

а) если перед прилагательным словосочетанием стоит слово из группы прилагательных, то оно дополнительно включается в это словосочетание.

б) если после прилагательного словосочетания стоит слово из группы существительных в родительном падеже, то оно дополнительно включается в это словосочетание.

2. Если после наречного или глагольного словосочетания стоит слово из группы существительных в родительном падеже, то оно дополнительно включается в это словосочетание.

3. Если после числительного или существительного словосочетания стоит слово из группы существительных в родительном падеже, то оно дополнительно включается в это словосочетание.

В примере (3.1) показан результат расширения двухсловных сочетаний ПС примера (3) в соответствии с изложенными выше правилами. Здесь в круглых скобках указан тип словосочетания, в квадратных – двусловные сочетания, в фигурных – расширенные по п.п..1 – 3 трёхсловные сочетания

(3.1) Пример:

Если Вам [необходимо активировать(2)] {[услугу передачи(5)] даных} {дляЪвашего [мобильного номера(1)]}.

Но благодаря {[разумному сочетанию(1)] лекарств} он [смог остановить(3)] [развитие болезни(5)] [вЪбольшинстве случаев(5)].

Тогда тарификация [Ваших звонков(1)] {[начинается сЪмомента(3)] соединения} [сЪтелефоном абонента(5)].

[Идеальным решением(1)] [проблемы членения(5)] {[такого рода(1)] предложений} наЪсинтагмы [былоЪбы использование(3)] {[комплекса правил(5)] разбора} [наЪсинтаксические компоненты(1)].

. В приведенном выше примере, несмотря на реализацию полной последовательности действий по разметке синтагм на словосочетания и по их расширению, остаются ещё отдельные слова, не вошедшие ни в одно из словосочетаний. Следующим, последним шагом, является дополнительное расширение словосочетаний путём включения в их состав слабоударных слов, таких, как многосложные предлоги, союзы и местоимения. Если же и после такого расширения словосочетаний остаются отдельные слова, то они определяются как частный случай однословных сочетаний.

Результат применения этого правила иллюстрируется примером (3.2).

(3.2) Пример:

{Если Вам [необходимо активировать(2)]} {[услугу передачи(5)] данных} {дляЪвашего [мобильного номера(1)]}.

{Но благодаря [разумному сочетанию(1)] лекарств} {он [смог остановить(3)]} [развитие болезни(5)] [вЪбольшинстве случаев(5)].

{Тогда [тарификация]} [Ваших звонков(1)] {[начинается сЪмомента(3)] соединения} [сЪтелефоном абонента(5)].

[Идеальным решением(1)] [проблемы членения(5)] {[такого рода(1)] предложений} [наЪсинтагмы] [былоЪбы использование(3)] {[комплекса правил(5)] разбора} [наЪсинтаксические компоненты(1)].

5. Расстановка слабых и сильных словесных ударений

Для двухсловных сочетаний в таблице 4.2 указано место предпочтительной установки сильного и слабого ударений, т.е. на первом или втором слове словосочетания. Такое распределение позиций слабого и сильного ударений не претендует, конечно, на универсальность. Оно характеризует среднестатистическую тенденцию для достаточно широкого набора различных текстов. При определённых условиях (индивидуальная манера чтения, стремление к определённой ритмической структуре и др.) знаки (+) (=) для данного словосочетания могут меняться местами, либо оба знака индицировать сильное ударение – (+) (+). Важную роль может играть также наличие некоторых индикаторов потенциальной «слабости» или «силы» какого-либо из слов в словосочетании. В частности, к индикатору «слабости» может быть отнесена принадлежность слова к группе потенциально слабоударных слов, таких, как многосложные предлоги и частицы, союзы и местоимения. К индикатору «силы» может быть отнесено наличие перед словом усилительной или отрицательной частицы.

050-Lobanov:_Layout 1 18.05.2008 15:45 Page 328

После применения указанных правил расстановки сильных и слабых ударений пример (3.2) перепишется в следующем виде.

(3.3) Пример:

{Е=сли Ва=м [необходи=мо активи+ровать(2)]} {[услу+гу переда=чи(5)] да=нных} {дляЪва=шего [моби=льного но+мера(1)]}.

{Но= благодаря= [разу=мному сочета+нию(1)] лека=рств} {о=н [смо=г останови+ть(3)]} [разви+тие боле=зни(5)] [вЪбольшинстве+ слу=чаев(5)].

{Тогда= [тарифика+ция]} [Ва=ших звонко+в(1)] {[начина+ется сЪмоме+нта(3)] соедине=ния} [сЪтелефо+ном абоне=нта(5)].

[Идеа=льным реше+нием(1)] [пробле+мы члене=ния(5)] {[тако=го ро+да(1)] предложе=ний} [наЪсинта+гмы] [бы+лоЪбы испо=льзование(3)] {[ко+мплекса пра=вил(5)] разбо=ра} [наЪсинтакси=ческие компоне+нты(1)].

На заключительном этапе целесообразно провести окончательную корректировку позиций сильных и слабых ударений с точки зрения приближения к оптимальной организации ритмической структуры синтагм. При этом уточняются ситуации, когда в ГрС имеется более одного слова со слабым ударением. Окончательная корректировка осуществляется, исходя из необходимости соблюдения следующих условий:

– в ГрС не должно быть двух следующих подряд слов со слабым ударением. В этом случае в одном из этих слов, например во втором, слабое ударение заменяется на сильное.

– в ГрС количество слов со слабым ударением не должно быть больше количества слов с сильным ударением. Например, последовательность (=) (+) (=) заменяется на последовательность (=) (+) (+).

Следует заметить, что приведенные здесь правила отражают лишь среднестатистические закономерности. Окончательные условия особой выделенности того или иного слова могли бы быть адекватно определены только в результате глубокого синтаксического и семантического анализа предложений, что в настоящий момент пока недостижимо.

После применения указанных правил корректировки пример (3.3) перепишется в следующем виде.

(3.4) Пример:

{Е=сли Ва+м [необходи=мо активи+ровать(2)]} {[услу+гу переда=чи(5)] да+нных} {дляЪва=шего [моби+льного но+мера(1)]}.

{Но= благодаря+ [разу=мному сочета+нию(1)] лека+рств} {о=н [смо+г останови+ть(3)]} [разви+тие боле=зни(5)] [вЪбольшинстве+ слу=чаев(5)].

{Тогда= [тарифика+ция]} [Ва=ших звонко+в(1)] {[начина+ется сЪмоме+нта(3)] соедине=ния} [сЪтелефо+ном абоне=нта(5)].

[Идеа=льным реше+нием(1)] [пробле+мы члене=ния(5)] {[тако=го ро+да(1)] предложе+ний} [наЪсинта+гмы] [бы+лоЪбы испо=льзование(3)] {[ко+мплекса пра=вил(5)] разбо+ра} [наЪсинтакси=ческие компоне+нты(1)].

6. Разметка ПЛС на акцентные единицы (АЕ) и синтаксические синтагмы

Разметка полученной в примере (3.4) последовательности слов на акцентные единицы осуществляется по следующим правилам:

1. Разметка на АЕ осуществляется раздельно для каждой ГСС.

2. Если в ГСС имеются слова со слабым ударением, то каждое из них объединяется в одну АЕ с сильноударным словом, стоящим слева или справа от него.

3. Оставшиеся слова с сильным ударением отмечаются как отдельные АЕ.

После применения указанных правил разметки на АЕ пример (3.4.) перепишется в следующем виде.

(3.5) Пример:

{(Е=сли Ва+м) (необходи=мо активи+ровать)} *[2]*
{(услу+гу переда=чи) (да+нных)} *[2]*
{(дляЪва=шего моби+льного) (но+мера)}. *[2]*
{(Но= благодаря+) (разу=мному сочета+нию) (лека+рств)} *[3]*
{(о=н смо+г) (останови+ть)} *[2]*
{(разви+тие боле=зни)} *[1]*
{(вЪбольшинстве+ слу=чаев)}. *[1]*
{(Тогда= тарифика+ция)} *[1]*
{(Ва=ших звонко+в)} *[1]*
{(начина+ется) (сЪмоме+нта соедине=ния)} *[2]*

050-Lobanov:_Layout 1 18.05.2008 15:45 Page 329

{(сЪтелефо+ном абоне=нта)}. *[1]*
{(Идеа=льным реше+нием)} *[1]*
{(пробле+мы члене=ния)} *[1]*
{(тако=го ро+да) (предложе+ний)} *[2]*
{(наЪсинта+гмы)} *[1]*
{(бы+лоЪбы испо=льзование)} *[1]*
{(ко+мплекса пра=вил) (разбо+ра)} *[2]*
{(наЪсинтакси=ческие компоне+нты)}. *[1]*

В примере (3.5) круглыми скобками отмечены полученные АЕ в каждом из ГрС, которые ограничены фигурными скобками и помещены на отдельных строках., причём справа от каждой строки указано количество АЕ в данной ГСС.

Как уже указывалось, основной смысл предварительного разбиения ПЛС на ФЛС и ГСС заключается в том, что теперь свобода разделения ПЛС на СС ограничивается, т.к. граница между СС не может находиться внутри ФЛС или ГСС. В простейшем случае границей каждой СС могут служить границы ГСС. В этом случае, как видно из примера (3.5), каждая СС будет включать различное количество АЕ: от 1-й до 3-х.

Если же требуемый стиль чтения предполагает, что СС должна включать по возможности не менее 2- х АЕ, то в этом случае получим схему членения, показанную на примере (3.6), где справа от каждой строки указано количество АЕ в данной СС.

(3.6) Пример:
{(Е=сли Ва+м) (необходи=мо активи+ровать)} *[2]*
{(услу+гу передa=чи) (да+нных)} *[2]*
{(дляЪва=шего моби+льного) (но+мера)} *[2]*
{(Но= благодаря+) (разу=мному сочета+нию) (лека+рств)} *[3]*
{(о=н смо+г) (останови+ть)} *[2]*
{(разви+тие боле=зни)} {(вЪбольшинстве+ слу=чаев)} *[2]*
{(Тогда= тарифика+ция)} {(Ва=ших звонко+в)} *[2]*
{(начина+ется) (сЪмоле+нта соедине=ния)} {(сЪтелефо+ном абоне=нта)} *[3]*
{(Идеа=льным реше+нием)} {(пробле+мы члене=ния)} *[2]*
{(тако=го ро+да) (предложе+ний)} {(наЪсинта+гмы)} *[3]*
{(бы+лоЪбы испо=льзование)} {(ко+мплекса пра=вил) (разбо+ра)} *[3]*
{(наЪсинтакси=ческие компоне+нты)} *[1]*

Заключение

Описанный алгоритм сегментации на синтагмы используется в составе системы синтеза речи по тексту «МультиФон» [7]. Образцы синтезированной речи будут продемонстрированы во время доклада.

Список литературы

1. Лобанов Б.М. и др. Алгоритмы синтеза просодических характеристик речи по тексту в системе ”Мультифон” // Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог’2007, М.: Издательский центр РГГУ, 2007. – С. 550-558.

2. Кривнова О.Ф. Фактор речевого дыхания в интонационно-паузальном членении речи // В кн: Лингвистическая полифония / Изд. «Языки славянских культур»– Москва, 2007 – С. 424-443.

3. Lobanov, B., Tsirulnik, L. Statistical study of speaker’s peculiarities of utterances into phrases segmentation // Speech Prosody: proceedings of the 3-rd International conference, Dresden, Germany, May 2–5, 2006. – Dresden, 2006. – V. 2. – P. 557–560.

4. Лобанов Б.М., Сизонов О.Г., Цирульник Л.И. Алгоритм интонационной разметки повествовательных предложений для синтеза речи по тексту / в наст. Сб. трудов Диалог’2008

5. Boguslavsky I., Karnevskaya E., Lobanov B. Generation of Intonation and Accentuation on the of Synthetic Speech on the Basis of Morpho-Syntactic Knowledge // Proc.of International Workshop «Integration of Language and Speech» – Moskow, 1995, pp. 11-28.

6. Валгина Н.С. Современный русский язык / http://www.hi-edu.ru/

7. Лобанов, Б.М. «МУЛЬТИФОН» - система персонализированного синтеза речи по тексту на славянских языках // В кн: Лингвистическая полифония / Изд. «Языки славянских культур»– Москва, 2007 – С. 849-866.

051-Lobanov:_Layout 1 18.05.2008 15:48 Page 330

«ИНТОКЛОНАТОР» – КОМПЬЮТЕРНАЯ СИСТЕМА КЛОНИРОВАНИЯ ПРОСОДИЧЕСКИХ ХАРАКТЕРИСТИК РЕЧИ

«INTOCLONATOR» – A COMPUTER SYSTEM FOR PROSODIC SPEECH PARAMETERS CLONING

Лобанов Б.М. *(Lobanov@newman.bas-net.by),* ***Цирульник Л.И.*** *(L.Tsirulnik@newman.bas-net.by),*
Сизонов О.Г. *(Osizonov@yahoo.co.uk)*
Объединённый институт проблем информатики НАН Беларуси, Минск, Беларусь

Описывается компьютерная система клонирования просодических характеристик речи – «ИнтоКлонатор», позволяющая автоматизировать процесс создания комплекса просодических портретов, необходимых для синтеза речи по произвольному тексту. Система предназначена для расширения инвентаря просодических портретов при синтезе персонализированной речи по текстам различных жанров.

Введение

Просодика играет важную роль как при восприятии смысла, так и при восприятии индивидуальности голоса и речи диктора. Поэтому просодическая модель, используемая при синтезе речи по тексту, должна адекватно отражать как языко-зависимые, так и дикторо-зависимые характеристики.

Существует достаточно большое число просодических моделей, предложенных для использования в системах синтеза речи по тексту. По методу представления интонации просодические модели можно разделить на следующие основные категории:

– автосегментная модель (АМ-модель) [1];
– суперпозиционная модель (СП-модель) [2];
– IPO-модель [3];
– непрерывная параметрическая модель (Tilt -модель)[4].

Алгоритмы и компьютерная система клонирования просодических параметров, рассматриваемые в данной работе, основаны на оригинальной модели представления интонации синтагмы последовательностью просодических Портретов Акцентных Единиц (ПАЕ). ПАЕ-модель была предложена более 20 лет назад [5] и успешно использовалось с тех пор в системах синтеза речи по тексту [6, 7].

В соответствии с ПАЕ-моделью, минимальной просодической единицей является Акцентная Единица (АЕ), состоящая из одного или более слов, и имеющая в своём составе только один полноударный слог. АЕ, в свою очередь, состоит из ядра (полноударный слог), предъядра (все фонемы, предшествующие полноударному слогу) и заядра (все фонемы за полноударным слогом). Главное предположение ПАЕ-модели состоит в том, что топологические свойства просодических параметров для определенного интонационного типа фразы не изменяются (или изменяются незначительно) с изменениями фонетического контекста и числа слогов в предъядре и заядре АЕ.

Для клонирования просодических характеристик речи диктора, прежде всего, записывается произносимый им специально подготовленный текст. Затем опытный фонетист анализирует полученную фонограмму и выделяет фонетические синтагмы (под синтагмой понимается самостоятельная в интонационном смысле часть фразы или вся фраза). Решение о наличии конца синтагмы принимается на основе ряда признаков, таких как: присутствие дыхательной паузы, комплексная реализация одного из возможных интонационных типов синтагмы, наличие определённой динамической структуры (контура силы звука) и определённой ритмической структуры (контура длительности звуков). При членении фонограммы на синтагмы во внимание принимается также присутствие знаков препинания в соответствующем ей тексте, а также некоторых других формальных признаков текста.

Каждая анализируемая синтагма автоматически размечается на акцентные единицы – АЕ. Затем осуществляется измерение просодических параметров для каждой АЕ: мелодики (значения частоты основного тона - F_0), динамики (значения амплитуды - A) и ритмики (значения длительности звуков - T), и формирование просодических портретов.

051-Lobanov:_Layout 1 18.05.2008 15:48 Page 331

1. Функциональная схема системы

Функциональная схема, входные и выходные данные, взаимодействие блоков системы представлены на рис 1.

Входные данные системы:

– предварительно обработанная фонограмма записи – набор речевых синтагм, каждая из которых хранится в виде оцифрованной звуковой волны в отдельном файле в формате *WAVE PCM*;

– предварительно обработанная стенограмма записи – набор текстовых файлов синтагм – с указанием интонационного типа и количества АЕ для каждой синтагмы;

– правила просодической маркировки синтезированного речевого сигнала на АЕ и элементы АЕ – предъядро, ядро и заядро.

Выходные данные системы:

БД просодических характеристик речи «клонируемого» диктора – набор мелодических, энергетических и ритмических портретов акцентных единиц.

Просодическая маркировка естественного РС. Текстовые файлы синтагм является входным данным блока синтеза и просодической маркировки синтезированного речевого сигнала (РС). В блоке синтеза осуществляется фонетическая и просодическая обработка текста, включающая преобразования «буква-фонема» и «фонема-аллофон», выбор звуковых волн аллофонов из акустической БД, их компиляция и маркировка синтезированного речевого сигнала на АЕ и элементы АЕ (ЭАЕ): предъядро, ядро и заядро. Результат обработки – просодически размеченная синтагма синтезированного речевого сигнала.

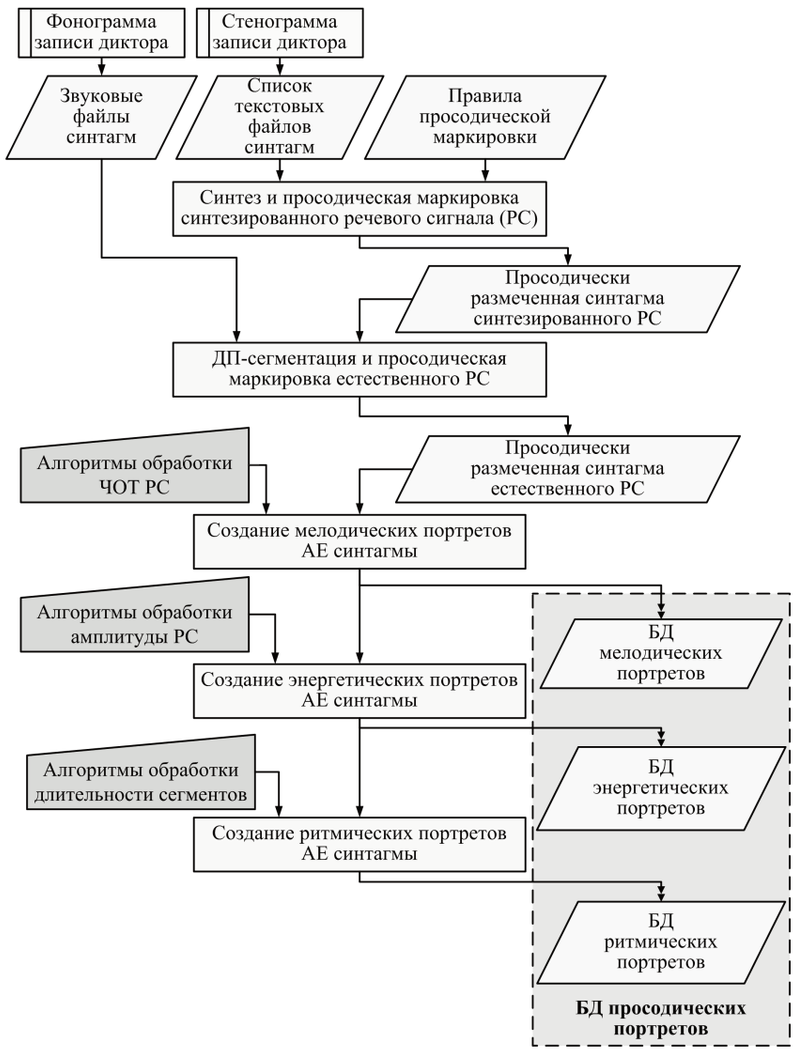


Рис. 1. Функциональная схема системы «ИнтоКлонатор»

051-Lobanov:_Layout 1 18.05.2008 15:48 Page 332

Каждая пара синтагм «просодически-размеченный синтезированный сигнал – естественный сигнал» поступает в блок ДП-сегментации и просодической маркировки естественного РС, в котором осуществляется разметка естественного сигнала на периоды основного тона (питчи), анализ акустических признаков естественного и синтезированного сигналов, их ДП-сопоставление и перенос маркеров границ аллофонов, АЕ и ЭАЕ с синтезированного на естественный РС. В системе реализована настройка параметров вычисления питчей естественного РС. Результатом работы блока является синтагма естественного РС, в которой расставляются метки питчей, аллофонов, а также предъядра, ядра и заядра для каждой АЕ. Для именования регионов приняты следующие обозначения: предъядро – preN, ядро – N, заядро – postN. По именам этих регионов определяются границы и длительности предъядра, заядра и ядра каждой акцентной единицы.

Пример сигнала синтагмы «*Машенька уснула*», размеченной на питчи, аллофоны, АЕ и ЭАЕ, показан на рис. 2. Синтагма состоит из двух АЕ: «Машенька» и «уснула». Ядром первой АЕ является аллофон A_{012}, ядром второй – аллофон U_{022}.

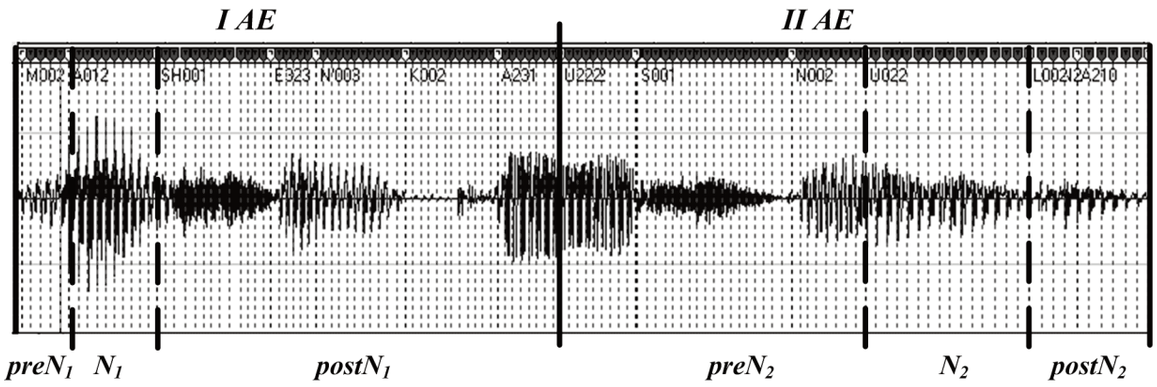


Рис. 2. Пример маркировки речевого сигнала синтагмы

2. Алгоритм создания мелодических портретов

С использованием информации о текущей длительности периодов речевого сигнала, задаваемой метками питчей, вычисляется исходный мелодический контур (ИМК) значений ЧОТ (F_0), при этом применяется процедура медианного сглаживания. Для каждого элемента АЕ – предъядра, ядра и заядра – равномерно выбирается пять точек ИМК, лежащих во временных пределах каждого элемента АЕ на участках, соответствующих аллофонам гласных и звонких согласных. При этом в ИМК не включаются точки, находящиеся в регионах аллофонов шумных согласных {*f, f', s, s', sh, sh', c, ch', h, h', p, p', t, t', k, k', b, b', d, d', g, g'*}. На участках шумных согласных реальные значения ЧОТ заменяются новыми значениями путём вычисления интерполяционной прямой от последней точки предшествующего региона звонкого аллофона к первой точке последующего региона звонкого аллофона. Пример обработки контура ЧОТ для синтагмы «*Машенька уснула*»показан на рис.3.

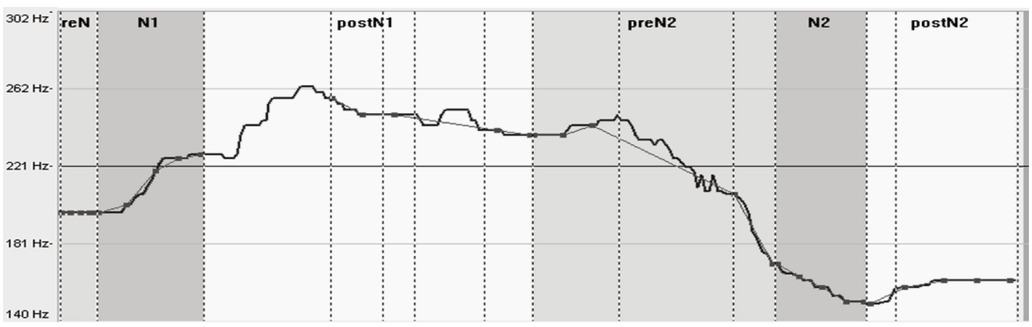


Рис. 3. Пример обработки мелодического контура синтагмы

На следующих шагах алгоритма осуществляется нормировка длительности сегментов preN, N, postN путём уравнивания длительности областей предъядра, ядра и заядра каждой АЕ, входящей в синтагму.

051-Lobanov:_Layout 1 18.05.2008 15:48 Page 333

Далее осуществляется нормировка контура ЧОТ. Для этого определяются минимальное – $F_{0\,\min}$ – и максимальное – $F_{0\,\max}$ – значения на всей исследуемой фонограмме. Нормированные значения ЧОТ вычисляется согласно формуле:

$$F_{0\,\text{norm}} = \frac{F_0 - F_{0\,min}}{F_{0\,max} - F_{0\,min}} \qquad (1)$$

Результататом описанных операций является создание последовательности нормированных мелодических портретов АЕ, составляющих синтагму (рис. 4).

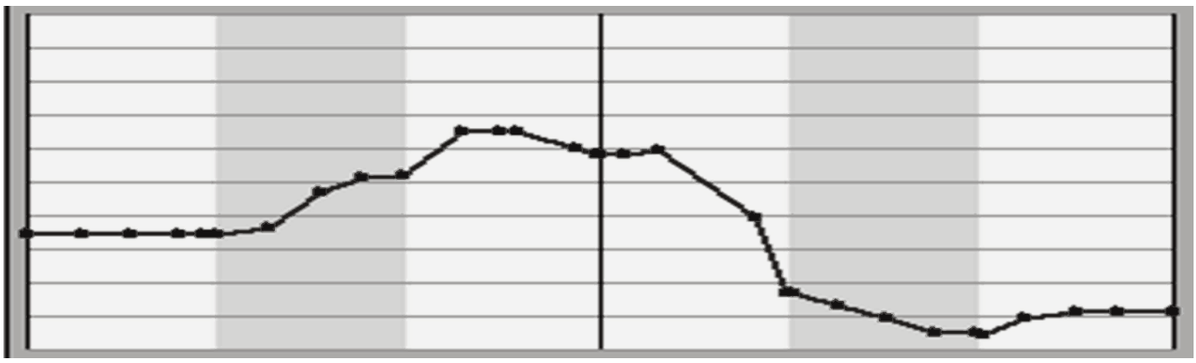


Рис. 4. Нормированный мелодический портрет двухакцентной синтагмы «Машенька уснула»

3. Алгоритмы создания энергетических и ритмических портретов

Для создания энергетического портрета синтагмы строится контур текущих значений энергии путём усреднения среднеквадратичного значения сигнала на интервале 15 миллисекунд с шагом 5 миллисекунд. На каждом из ядер АЕ синтагмы выбирается максимальное значение текущей энергии – $A_{i\,max}$. Графическое построение контура производится по следующему правилу. От левой границы сигнала до правой границы первого ядра строится горизонталь на уровне значения этого ядра $A_{1\,max}$. Далее от правой границы первого ядра до точки со значением амплитуды второго ядра синтагмы $A_{2\,max}$ строится прямая, а от неё до правой границы этого же ядра строится горизонталь. И так далее, до последнего ядра, от правой границы которого проводится горизонталь до конца сигнала.

Пример обработки энергетического контура для синтагмы «*Машенька уснула*»показан на рис. 5.

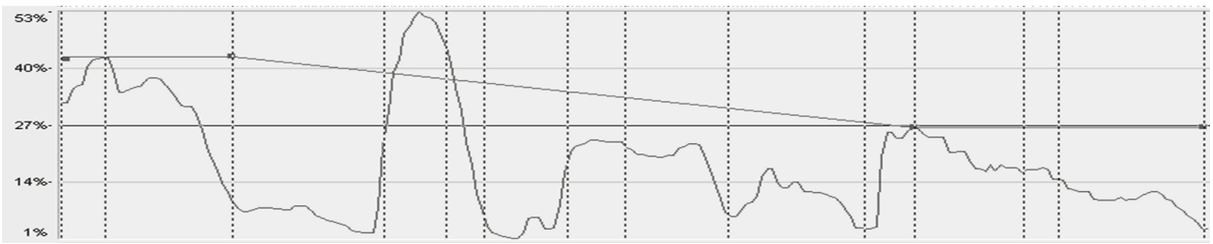


Рис. 5. Пример обработки энергетического контура синтагмы

Далее производится нормировка энергетического контура контура. Нормировка по длительности сегментов preN, N, postN осуществляется, как и в предыдущем случае, путём уравнивания длительности областей предъядра, ядра и заядра каждой АЕ, входящей в синтагму. Нормировка энергетических уровней осуществляется путём деления полученного энергетического контура на величину наибольшего значения $A_{i\,max}$, найденного на всей исследуемой фонограмме.

Результататом описанных операций является создание нормированного энергетического портрета синтагмы (рис. 6).

051-Lobanov:_Layout 1 18.05.2008 15:48 Page 334

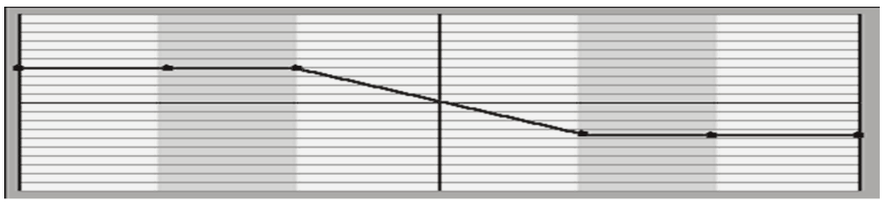


Рис. 6. Нормированный энергетический портрет двухакцентной синтагмы «Машенька уснула»

Для создания ритмического портрета осуществляются следующие операции. Вычисляются длительности ядер АЕ, входящих в синтагму – T_{N1}, T_{N2}, T_{N3}, …. Определяется максимальная из длительностей ядер в синтагме и осуществляется вычисление нормированных ритмических коэффициентов изменения длительности ядер в синтагме относительно ядра с максимальной длительностью. Ритмический коэффициент *i*-ой АЕ R_i вычисляется в соответствии с формулой

$$R_i = \frac{T_{Ni}}{T_{Ni\,max}} \qquad (2)$$

где T_{Ni} – длительность ядра *i*-й АЕ синтагмы, $T_{Ni\,max}$ – максимальная из длительностей ядер в синтагме.

Результататом описанных операций является создание нормированного ритмического портрета синтагмы (рис. 7). Нижний участок рисунка показывает изменённые под действием ритмического фактора относительные длительности ядер первой и второй АЕ синтагмы.

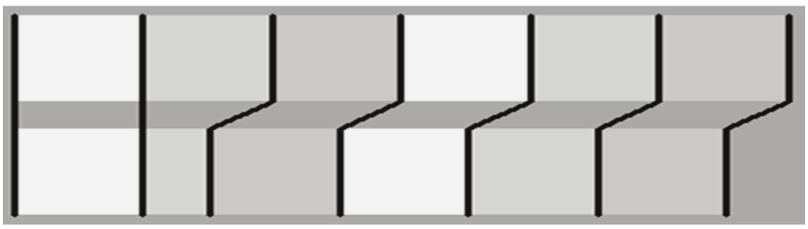


Рис. 7. Нормированный ритмический портрет двухакцентной синтагмы

4. Пользовательский интерфейс системы «ИнтоКлонатор»

Пользовательский интерфейс системы «ИнтоКлонатор» (рис. 8) включает следующие блоки:
– окно отображения осциллограммы речевого сигнала (РС);
– окно отображения графика нормированной амплитуды (A_{norm}) сигнала;
– окно отображения графика нормированной ЧОТ ($F_{0\,norm}$) сигнала;
– диалоговые окна настроек параметров системы.

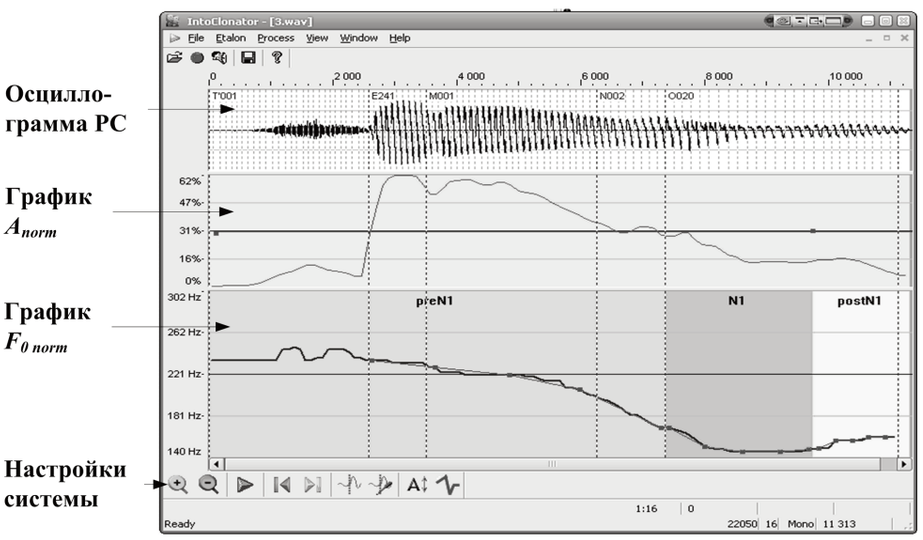


Рис. 8. Общий вид пользовательского интерфейса системы «ИнтоКлонатор»

051-Lobanov:_Layout 1 18.05.2008 15:48 Page 335

На осциллограмме РС (рис. 9) указаны границы периодов основного тона и аллофонов, а также имена аллофонов. В системе реализовано масштабирование отображения и прослушивание выделенного фрагмента РС.

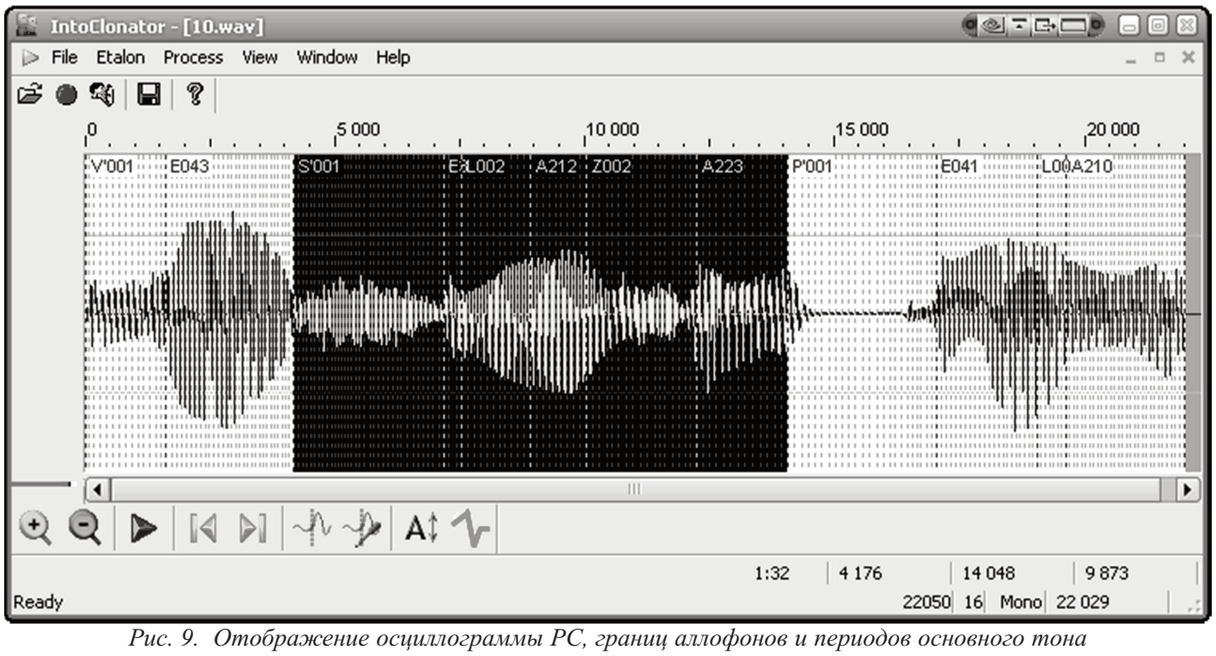


Рис. 9. Отображение осциллограммы РС, границ аллофонов и периодов основного тона

Нормированные амплитуда и ЧОТ сигнала (рис 10) вычисляются в соответствии с задаваемым диапазоном A_{min}, A_{max} и $F_{0\,min}$, $F_{0\,max}$. На графиках отображаются границы АЕ синтагмы и предъядра, ядра и заядра каждой АЕ, а также вычисленные динамический и мелодический портреты.

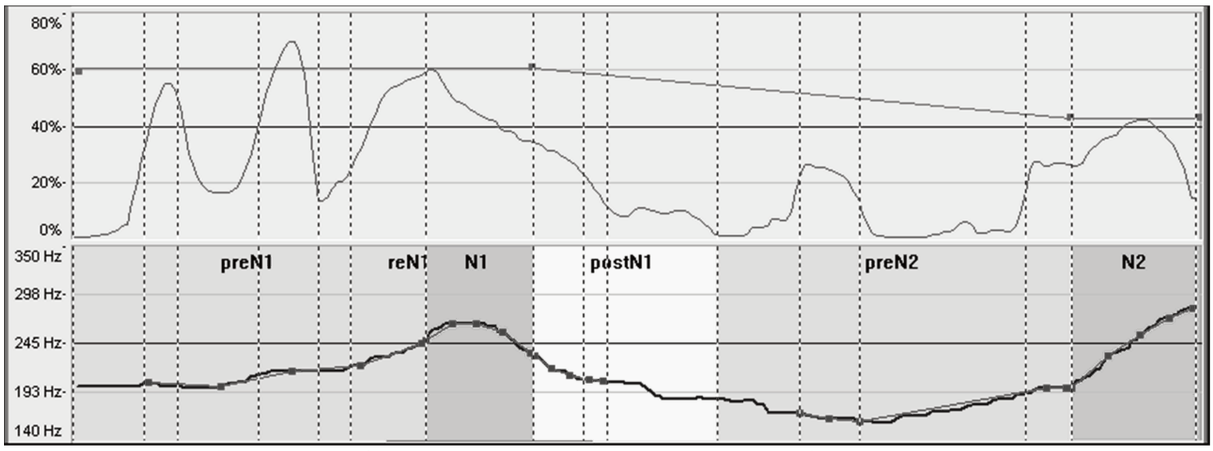


Рис. 10. Отображение графиков A_{norm}, $F_{0\,norm}$, динамического и мелодического портретов синтагмы

Настройки параметров системы реализованы в следующих диалоговых окнах:

– диалог настроек параметров вычисления ЧОТ;

– диалог настроек параметров сегментации РС,

– диалог установки диапазона амплитуды и ЧОТ;

Настройки блока вычисления ЧОТ позволяют устанавливать параметры вычисления спектральных характеристик и параметры определения вокализованных участков сигнала.

Настройки блока сегментации позволяют устанавливать параметры ДП-сопоставления естественного и синтезированного РС.

Настройки диапазона амплитуды и ЧОТ позволяют указывать значения A_{min}, A_{max} и $F_{0\,min}$, $F_{0\,max}$, которые должны быть определены заранее для набора речевых синтагм, обрабатываемых системой.

051-Lobanov:_Layout 1 18.05.2008 15:48 Page 336

5. Результаты практического использования системы «ИнтоКлонатор»

Система «ИнтоКлонатор» работает на базе специально разработанного текстового корпуса, включающего «мини-текст» для создания основного набора просодических портретов и «макси-тексты» для создания расширенного набора просодических портретов русской речи. С использованием системы «ИнтоКлонатор» создана БД просодических портретов для 1-й версии системы синтеза русской речи по тексту «МультиФон -1», включающая мелодические, динамические и ритмические портреты для следующих интонационных типов.

Для повествовательных предложений.

Синтагмы с интонацией незавершённости, которые образуются в следующих ситуациях:

1) С1, если «И»;

2) С2, если «ИЛИ»;

3) С3, если «,» и не С7 – С11 при условии, что «,» встретилась в тексте впервые или в 4-й, 7-й,… раз подряд;

4) С3_1, если «,» и не С7 – С11 при условии, что «,» встретилась в тексте во 2-й, 5-й, 8-й,… раз подряд;

5) С3_2, если «,» и не С7 – С11 при условии, что «,» встретилась в тексте в 3-й, 6-й, 9-й раз подряд;

6) С4, если «-»;

7) С5, если «(»;

8) С6, если «, - »;

9) С7, если «,» и союз сочинительный;

10) С8, если «,» и союз вопросительно- подчинительный;

11) С9, если «,» и союз подчинительный;

12) С10, если «,» и причастие;

13) С11, если «,» и деепричастие;

14) С01, если первая, третья, пятая и т.д. синтаксическая синтагма;

15) С02, если вторая, четвёртая и т.д. синтаксическая синтагма.

Синтагмы с интонацией завершённости, которые образуется в следующих ситуациях:

16) Р1, если «:»;

17) Р2, если «);

18) Р3, если «;»;

19) Р4, если «.» при условии, что «.» встретилась в тексте в1-й или 4-й, 7-й и т.д. раз подряд;

20) Р4_1, если «.» при условии, что «.» встретилась в тексте во 2-й, 5-й, 8-й и т.д. раз подряд;

21) Р4_2, если «.» что «.» встретилась в тексте в 3-й, 6-й, 9-й … раз подряд;

22) Р5, если «…»;

23) Р5, если «.» и конец абзаца;

24) Р6, если «.» и конец текста;

25) Р7, если «.» и в начале союз сочинительный после (,);

26) Р8, если «.» и в начале союз вопросительно- подчинительныq после (,);

27) Р9, если «.» и в начале союз подчинительный после (,);

28) Р10, если «.» и в начале причастие после (,);

29) Р11, если «.» и в начале деепричастие после (,).

Для вопросительных предложений:

30) Q1, если в составе синтагмы имеется вопросительное слово и если в вопросительном предложении оказалась только одна синтагма;

31) Q2, если в составе синтагмы отсутствует вопросительное слово и если в вопросительном предложении оказалась только одна синтагма;

32) Q1-1, если в составе синтагмы имеется вопросительное слово и если в вопросительном предложении более, чем одна синтагма;

33) Q2-1, если в составе синтагмы отсутствует вопросительное слово и если в вопросительном предложении более, чем одна синтагма.

Для восклицательных предложений.

34) Е1, если в составе синтагмы имеется междометие и если в восклицательном предложении оказалась только одна синтагма;

35) Е2, если в составе синтагмы отсутствует междометие и если в восклицательном предложении оказалась только одна синтагма;

36) Е1_1, если в составе синтагмы имеется междометие и если в восклицательном предложении более, чем одна синтагма;

051-Lobanov:_Layout 1 18.05.2008 15:48 Page 337

37) E2_1, если в составе синтагмы отсутствует междометие и если в восклицательном предложении более, чем одна синтагма.

Итого с использованием системы «ИнтоКлонатор» созданы просодические портреты синтагм 37-ми интонационных типов. При этом допускались 4 возможных варианта синтагм, состоящих из одной, двух, трёх и четырёх акцентных единиц. Таким образом, были созданы 37*4 = 148 просодических портретов.

Для анализа персональных особенностей реализации мелодических портретов АЕ четырёх интонационных типов: незавершённость, завершённость, восклицание и вопрос, были проведены исследования в соответствии со следующей методикой. Два профессиональных диктора радио (Олег и Светлана) и три непрофессиональных (Борис, Елена, Лилия) зачитали один и тот же отрывок художественного текста. Затем в фонограммах записей каждого диктора были выделены одни и те же участки речи, на которых ими были реализованы указанные четыре интонационных типа, и на основании анализа контуров F_0 построены ПАЕ в соответствии с разработанной методикой. На рис. 11 представлены полученные мелодические портреты конечных АЕ для четырёх интонационных типов 5-ти дикторов.

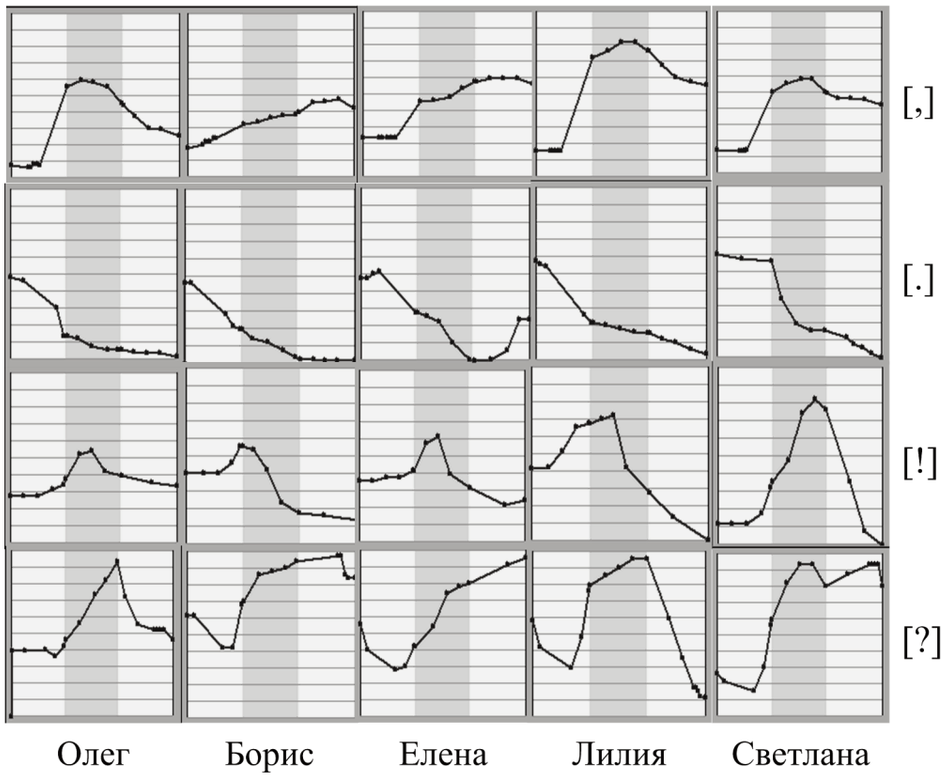


Рис. 11. Мелодические портреты конечной АЕ для четырёх интонационных типов 5-ти дикторов

Как видно из рис. 11, полученные мелодические ПАЕ имеют достаточно ярко выраженные персональные особенности. Причём в наибольшей степени индивидуальные дикторские различия в ПАЕ проявляются на предъядре и заядре, в то время как на ядерных участках они менее значительны. В целом же, однако, сохраняется рисунок портретов, характерный для каждого исследуемого интонационного типа.

Заключение

Разработанная система клонирования просодических характеристик речи позволила во много раз сократить трудоёмкость и время, необходимые для создания комплекса просодических портретов для синтеза речи по произвольному тексту. Система «ИнтоКлонатор» находит применение как для нужд дальнейшего расширения инвентаря ПАЕ при синтезе речи по текстам различных жанров, так и при создании персонализированных БД ПАЕ.

Доклад будет проиллюстрирован образцами синтезированной речи, просодические характеристики которых создавались с помощью системы «ИнтоКлонатор».

051-Lobanov:_Layout 1 18.05.2008 15:48 Page 338

Список литературы

1. Silverman, K. et al. TOBI: a standard for labelling English prosody. ICSLP: 867-870, 1992.

2. Fujisaki, H. Prosody, Models, and Spontaneous Speech. Computing Prosody, Springer-Verlag: 27–42, 1996.

3. De Pijper, J. Modelling British English Intonation. Foris, Dordrecht: 1983.

4. Taylor, P. Analysis and synthesis of intonation using the Tilt model. J. Acoust. Soc. of America: 2000.

5. Lobanov B., The Phonemophon Text-to-Speech System. 11-th ICPhS, Tallin: 1987, 61-64.

6. Lobanov B., Tsirulnik L., Zhadinets D., Karnevskaya E. Language- and Speaker Specific Implementation of Intonation Contours in Multilingual TTS Synthesis. Speech Prosody, Dresden: 2006, v. 2, 553-556.

7. Лобанов, Б.М. «МУЛЬТИФОН» - система персонализированного синтеза речи по тексту на славянских языках // В кн: Лингвистическая полифония / Изд. «Языки славянских культур»– Москва, 2007 – С. 849-866.

052-Luckashevich:_Layout 1 18.05.2008 15:49 Page 339

ОТБОР СЛОВОСОЧЕТАНИЙ ДЛЯ СЛОВАРЯ СИСТЕМЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ

AUTOMATED ANALYSIS OF MULTIWORD EXPRESSIONS FOR COMPUTATIONAL DICTIONARIES

Лукашевич Н.В. *(louk@mail.cir.ru),* ***Добров Б.В.*** *(dobroff@mai.cir.ru),* ***Чуйко Д.С.*** *(Dasha_C@mail.ru)*
Научно-исследовательский вычислительный центр МГУ им. М.В. Ломоносова;
АНО Центр информационных исследований

В статье описываются принципы работы системы автоматизированного анализа словосочетаний, помогающей экспертам обнаруживать особенности словосочетаний на основе их компонентой структуры. Система получает на вход список словосочетаний, автоматически извлеченный по текстовой коллекции, и в качестве словарной базы анализа использует лингвистический ресурс тезаурусного типа. Эксперименты с системой мы проводим в рамках построения Онтологии по естественным наукам и технологиям - ОЕНТ.

1. Введение

Известно, что компьютерные словари, словари систем автоматической обработки текстов должны включать описания не только значений отдельных слов, но и разного рода устойчивых словосочетаний, терминов. Поэтому важным является ответ на два вопроса. Во-первых, каковы должны быть принципы отбора словосочетаний в словарь. Во-вторых, насколько процесс отбора словосочетаний может быть автоматизирован.

С одной стороны, чем больше в компьютерных словарях описано словосочетаний, тем меньше проблем с разрешением многозначности отдельных слов, тем больше будет зафиксировано специфических случаев сочетаемости. Так, в работах Большакова И.А. [2] предлагается набирать в специальную базу Кросс-лексика все встретившиеся словосочетания. С другой стороны, большие объемы неструктурированного материала трудно использовать в приложениях компьютерной обработки текстов.

Традиционным подходом является описание в компьютерных словарях семантически связанных словосочетаний (идиом, фразеологизмов), которые демонстрируют какие-либо отклонения в синтаксическом и/или семантическом поведении [1, 7]. Спектр таких устойчивых словосочетаний очень широк: от жестко фиксированных словосочетаний, которые могут рассматриваться как «слово с пробелами», до словосочетаний, которые подчиняются практически всем синтаксическим и семантическим правилам языка лишь за некоторым исключением. В последнем случае сразу обнаружить такую особенность может быть весьма сложно.

В работе Sag et.al. [12] обсуждается важный вид словосочетаний, называемых авторами институциональными выражениями. Для таких выражений характерно то, что по большей части эти выражения выглядят как свободные словосочетания, однако их компоненты не всегда могут быть заменены синонимами. Кроме того, частотность такого словосочетания очень высока по сравнению с теми словосочетаниями, которые образованы заменой слов-компонентов на синонимы. Примером таких словосочетаний является словосочетание *phone booth (телефонная будка)*. Так, и в русском, и в английском языке попытка замены слова *booth (будка)* на другие слова, например, кабина, приводит к многократному снижению частотности употребления.

Задача автоматизации отбора устойчивых словосочетаний, терминов в словарь далека от полного решения. Предложено множество методов и алгоритмов извлечения устойчивых словосочетаний, терминов из текстов (см. например, работы [3, 6, 8] и указанную в них литературу). Стандартные методы приводят, по большому счету, к одному и тому же результату. Обычно в результате программы отбора словосочетаний порождается список, упорядоченный по весу в соответствии с заложенной моделью. Верхняя часть такого списка наполнена терминами, устойчивыми словосочетаниями, которые необходимо включать в словарь прикладной системы.

Далее процент очевидно нужных для описания словосочетаний резко снижается, и наибольшую долю начинают составлять словосочетания, для которых очень трудно решить, нужно ли их описывать в словаре - для этого требуется серьезный дополнительный анализ. Поэтому методы автоматического извлечения словосочетаний терминологических словосочетаний достаточно трудно оценивать [6]. Привлеченные эксперты

052-Luckashevich:_Layout 1 18.05.2008 15:49 Page 340

часто дают очень противоречивые оценки [3]. При работе с большими предметными областями и корпусами даже лучшие методы извлечения терминологических словосочетаний показывают падение процента терминов с 90% на первой сотне до 60% на третьей тысяче списка извлеченных терминологических словосочетаний [6].

В реальной работе по формированию словарей верхняя часть (первые несколько сотен единиц) полученного списка достаточно быстро обрабатывается экспертами. Актуальной задачей является снижение трудоемкости работы экспертов при обработке сравнительно малочастотных словосочетаний.

В данной статье мы рассмотрим принципы выявления словосочетаний, необходимых для внесения в словарь компьютерной системы обработки текстов, возможные способы их формализации для автоматического выявления таких словосочетаний. Также мы опишем разрабатываемую нами систему автоматизированного отбора словосочетаний, которая должна помогать лингвистам, терминологам, экспертам выявлять особенности извлеченных словосочетаний и облегчать принятие решений по поводу их включения/невключения в словарь.

2. Критерии внесения словосочетаний в словарь

Рассмотрим, какие факторы можно учитывать, принимая решения о внесении словосочетаний в словарь.

Разработчики информационно-поисковых тезаурусов традиционно выделяют особое внимание отбору многословных терминов для включения в тезаурус.

Так, в стандартах по разработке информационно-поисковых тезаурусов [4, 9] указывает, что допускается включать словосочетания в словник, если в качестве опорного слова они содержат существительное и если выполнено одно из следующих условий:

значение словосочетания не выводится из значений его компонентов (*черный ящик*);

- хотя бы один из компонентов словосочетания не употребляется в составе других сочетаний или употребляется всегда в другом смысле (*торговля на вынос*);
- для данного словосочетания в словнике ИПТ существуют полные синонимы, например, *высоколиквидные акции – голубые фишки* ;
- отдельные слова словосочетания имеют слишком широкое значение;
- имеется общепринятая аббревиатура.

В работах [9, 11, 13] обсуждается совокупность принципов, которые могут служить (в сочетании) основанием для внесения словосочетания в компьютерный словарь:

- высокая частотность;
- высокая степень ассоциации, то есть более частое употребление друг с другом , чем с другими словами;
- синонимичность лексической единице (например, отдельному слову);
- значительная многозначность компонентов (*состояние дел, повестка дня*);
- словосочетание обозначает тип объекта, например, *телефонная будка, письменный стол.*

Таким образом, мы видим, что различные авторы предлагают различные критерии и соображения для включения многословных конструкций в словари компьютерных систем, что значительно затрудняет принятие решения в конкретных случаях.

3. Методы автоматического извлечения устойчивых словосочетаний

Существующие методы автоматического извлечения устойчивых словосочетаний, терминов обычно используют некоторое сочетание следующих факторов:

- частотные характеристики словосочетания (частотность по коллекции, взаимная ассоциация, вхождение в объемлющие словосочетания и т.п.);
- синтаксические ограничения: извлекаются словосочетания заданной синтаксической структуры: именные группы, глагольные группы;
- лексические фильтры, например, не извлекаются словосочетания, включающие географические названия.

Между тем есть еще ряд факторов, которые можно использовать, на следующих этапах отбора устойчивых словосочетаний.

Во-первых, часто словосочетания не извлекаются «с нуля», обычно имеется некоторый исходный ресурс, содержащий значения отдельных слов, уже включающий некоторый набор словосочетаний, и таким образом извлеченные словосочетания, можно анализировать относительно имеющегося словаря, тезауруса и т.п.

Во-вторых, извлеченные словосочетания можно сравнивать друг с другом и находить какие-либо особенности или сходство между ними.

052-Luckashevich:_Layout 1 18.05.2008 15:49 Page 341

В-третьих, можно изучать особенности словосочетания, проверяя употребление словосочетаний (исходных или специальным образом порожденных) в Интернет.

Например, в работе [12] предлагается использовать для извлечения устойчивых словосочетаний синонимы, описанные в тезаурусе WordNet. Поскольку одним из частых свойств семантически связанных словосочетаний является ограничение на замену одного из слов словосочетания синонимом, то предлагается исследовать сочетания синонимов с одними и теми же словами по корпусу, затем перепроверять в Интернет. Если разница частотностей таких словосочетаний значительна, то можно предлагать частотное словосочетание как устойчивое. Например, сравнивая употребление слов-синонимов *baggage* и *luggage* в сочетаниях с различными словами, можно обнаружить, что только *baggage* употребляется с таким прилагательным как *emotional*. Таким образом, можно предположить, что словосочетание *emotional baggage* является устойчивым. Однако в реальности ситуация осложняется тем, что у всех слов в составе исследуемых словосочетаний может быть по несколько значений, и встретившиеся словосочетания могут включать слова в разных значениях.

В работе [14] описан эксперимент по пополнению WordNet новыми словами и словосочетаниями, извлеченными из текстов путем встраивания в иерархии WordNet. Процедура реализуется за счет автоматического сопоставления сочетаемости слов и словосочетаний из WordNet с сочетаемостью неизвестных выражений. Проведенная нами проверка представленных авторами работы результатов показала, что большинство дополненных слов и словосочетаний представляют собой имена конкретных сущностей, а устойчивых словосочетаний и терминов очень мало. Отметим, что для сборки именованных объектов существуют специализированные эффективные методы.

Для исследования возможностей перечисленных выше дополнительных факторов в распознавании устойчивых словосочетаний, терминов была начата разработка автоматизированной системы анализа словосочетаний АРМ «Словосочетание».

Входом для системы анализа словосочетаний служат списки словосочетаний, автоматически извлеченных из текстовой коллекции на основе устоявшихся технологий извлечения [5, 8]. Словосочетания снабжены данными об их статистических характеристиках (частотности, взаимной ассоциации и т.п.). Состав словосочетания описывается набором слов в словарной форме, порожденных автоматически.

Словарной базой анализа словосочетаний является лингвистический ресурс тезаурусного типа, в котором синонимия слов описывается через отнесение к одной и той же единице тезауруса (дескриптору, синсету, понятию, концепту – далее концепт), разные значения слова отнесены к разным единицам тезауруса, а отношения между единицами тезауруса описаны в виде формализованных отношений. Таким образом, может использоваться тезаурус типа WordNet. Мы предполагаем использовать тезаурусы типа Тезаурус русского языка РуТез [5].

Словосочетания, по которым принято решение об их необходимости внесения в словарь, целесообразно заносить также в словарь тезаурусного вида. Таким образом, уже принятые решения смогут указывать влияние на дальнейший анализ словосочетаний.

4. Методы дополнительного анализа словосочетаний на примере пополнения Онтологии по естественным наукам и технологиям ОЕНТ

В настоящее время реализация и тестирование системы анализа словосочетаний проводится в рамках работ по развитию онтологии по естественным наукам и технологиям – ОЕНТ [5]. Онтология ОЕНТ представляет собой лингвистический ресурс для автоматической обработки текстов и содержит терминологию таких наук как математика, физика, химия, геология, биология, мы предполагаем ее бесплатное распространение для некоммерческого применения.

Начав работы над этой онтологией в 2004, мы собрали текстовые коллекции по разным естественным наукам (от 3000 до 8000 документов, от 50 до 90 Мб по каждой из наук), автоматически извлекли из них терминологические словосочетания (более 600 тысяч словосочетаний) и наиболее частотные их них (60 тысяч словосочетаний) стали одним из источников терминологии онтологии ОЕНТ.

В настоящее время величина онтологии ОЕНТ составляет 50 тысяч концептов, 135 тысяч терминов. Многие источники терминов (энциклопедии, учебники, учебные материалы) уже использованы для терминологического пополнения ОЕНТ, и существенным вопросом становится проверка полноты покрытия онтологии, пополнение относительно новыми терминами (еще не содержащимися в энциклопедических ресурсах, или не выделенными в заголовки энциклопедических статей). Поэтому мы снова обратились к исходным спискам словосочетаний как ресурсу пополнения онтологии.

Число оставшихся необработанными словосочетаний слишком велико для подробной работы. Поэтому для дальнейшего анализа извлеченных словосочетаний необходимо применение автоматизированных методов.

052-Luckashevich:_Layout 1 18.05.2008 15:49 Page 342

Рассмотрим, с чем сталкивается эксперт при анализе такого рода извлеченных списков словосочетаний, на следующем фрагменте списка словосочетаний, собранных для ОЕНТ (упорядочен по алфавиту):

вторичное зеркало
дальнейшее продвижение
изотропный источник
кристаллический агрегат
лист растения
нервное волокно
оператор проекции момента импульса
основное уравнение динамики вращательного движения
относительная внешняя система координат
параметры атомов
переходное состояние
порошкообразный алюминий
промежуточное ядро
псевдоскалярный мезон
радиоактивный ряд
степень доктора
тепловое давление
энергетическая единица
элементарная ячейка обратной решетки

Просматривая этот список, можно достаточно легко увидеть, что словосочетание *дальнейшее продвижение* – это явно не термин, поскольку не принадлежит к какой-либо конкретной области, а *нервное волокно* – это явный термин, поскольку обнаруживается в качестве заголовков статей многих словарей и энциклопедий. Для принятия решений по остальным словосочетаниям необходимо проведение дополнительных проверок.

Можно выделить следующие основные методы автоматизированного анализа новых словосочетаний.

1) Имея базовый тезаурус, можно выявлять словосочетания, которые близки по составу к словам или словосочетаниям базового тезауруса, к другим рассматриваемым словосочетаниям, то есть структурных синонимов, например, *порошкообразный алюминий – алюминиевый порошок, лист растения – листовое растение, учебный объект – учебный предмет, энергетическая единица – единица энергии*.

На первый взгляд представляется, что выявление структурных синонимов словосочетания подчеркивает тот факт, что словосочетание не является устойчивым, и не требуется описывать его в словаре. Однако приведенные выше примеры показывают, что ситуация значительно сложнее:

- для уже занесенного в словарь словосочетания может быть найден неочевидный синоним;
- может быть выявлено, что структурные синонимы, на самом деле, не являются смысловыми синонимами, так *листовое растение* – это *не растение с листьями*, а словосочетания *учебный объект* и *учебный предмет* обозначают разные сущности;
- нахождение группы структурных синонимов в извлеченных из текстов словосочетаниях обычно означает важность обозначаемой ими темы, поможет эксперту обнаружить пропущенные термины и принять решение, например, *агрегат кристаллов – кристаллический агрегат.*

При анализе структуры словосочетания могут быть также быть обнаружены словосочетания, являющиеся синонимами отдельных слов. Так, словосочетание *лист растения* является синонимом одного из значений слова *лист.* Такие, как бы тавтологичные, словосочетания могут быть обнаружены по структуре тезауруса (концепты слов словосочетания связаны между собой тезаурусными отношениями) или по толкованиям (одно слово из словосочетания встречается в толковании другого слова словосочетания**:** *лист1 – Орган воздушного питания и газообмена у растений*...).

2) На основе структуры анализа структуры словосочетаний важно выявлять словосочетания, повышающие структурированность тезауруса, представляющих собой полезные обобщения уже описанных в ресурсе сущностей. В приведенном списке таким словосочетанием является словосочетание *параметры атомов.* Введенный концепт с тем же названием будет обобщающим для таких концептов как *ЗАРЯД АТОМА, РАЗМЕР АТОМА, АТОМНЫЙ ВЕС* и др.

052-Luckashevich:_Layout 1 18.05.2008 15:49 Page 343

3) Если в ресурсе уже сформированы некоторые семантические классы слов в виде иерархической системы, то важно проверять соответствие семантических классов зависимых слов в словосочетаниях с одним и тем же главным словом. Так, если в ряду извлеченных словосочетаний *степень гидролиза, степень диссоциации, степень дифференциации, степень диффузности*, представляющих собой сочетание слова *степень* со словами, обозначающими процессы и свойства, встретится словосочетание *степень доктора*, то это словосочетание важно предъявить эксперту, поскольку значения слова *доктор* относятся к семантическому классу =человек=. Такое непохожее поведение зависимых слов в конструкциях с одними и теми же главными словами оценивается специальными весами.

4) На современном уровне развития интернет-технологий необходимым является проведение разного рода проверок употребления словосочетания в Интернет. Через глобальные поисковые сервисы можно проверить частотность употребления словосочетания (иногда относительно высокая частотность словосочетания в коллекции может оказаться случайностью [3]), а также возможно проверять, не имеет ли словосочетание каких-либо отношений, не вытекающих из его структуры.

Так, для словосочетания *вторичное зеркало* было выявлено частотное совместное употребление в текстах Интернет со словами *телескоп* и *рефлектор*, что объясняется тем, что вторичное зеркало является частью устройства телескопа-рефлектора, то есть концепт *ВТОРИЧНОЕ ЗЕРКАЛО* должен быть включен в онтологию ОЕНТ.

Анализ частот совместной встречаемости в поисковой выдаче слов со словосочетанием *радиоактивный ряд* показывает высокую частоту встречаемости слова *семейство*. Экспертная проверка показала, что словосочетание *радиоактивный ряд* является важным термином *(цепочка радиоактивных превращений),* а словосочетание *радиоактивное семейство* приводится как синоним данного термина.

Таким образом, описанные выше процедуры позволяют обратить внимание экспертов на значимые особенности словосочетаний и легче принимать решения об их включении/невключении в онтологию.

Для первого эксперимента по автоматизации дополнительных проверок извлеченных из текстов словосочетаний мы выделили следующие по частотности 20 тысяч словосочетаний (то есть с 60-й по 80-ю тысячу извлеченных словосочетаний), и применили к ним и к онтологии ОЕНТ, как словарной базе, описанные выше процедуры.

В результате было выделено около 1000 словосочетаний (5%) с особенными характеристиками. Эти словосочетания были предъявлены экспертам вместе с указанием типа найденной особенности, что позволяет относительно легко принимать решение. На текущий момент работа системы уже привела к пополнению онтологии ОЕНТ 150 терминами-синонимами, было введено 10 новых концептов.

Мы планируем расширить анализируемую базу словосочетаний до сотен тысяч, одновременно увеличить количество фильтрующих процедур, учитывающих иерархию тезаурусных отношений и статистику встречаемости словосочетаний и их структурных компонент в Интернет.

Заключение

При создании компьютерных словарей одной из трудоемких задач является обнаружение устойчивых словосочетаний и терминов для описания их в словаре. Особую сложность представляет собой анализ большого количества средне- и малочастотных словосочетаний широкой предметной области. Проблема связана с тем, что имеется множество разных видов словосочетаний, имеющих синтаксические и семантические особенности употребления, которые не следуют из состава этих словосочетаний.

В статье мы описали принципы реализации системы автоматизированного анализа словосочетаний, помогающей экспертам обнаруживать особенности словосочетаний на основе их компонентой структуры. Система получает на вход список словосочетаний, автоматически извлеченный по текстовой коллекции, и в качестве словарной базы анализа использует лингвистический ресурс тезаурусного типа. Первые эксперименты с системой мы проводим в рамках построения Онтологии по естественным наукам и технологиям - ОЕНТ.

Список литературы

1. Баранов А.Н., Добровольский Д.О. К универсальному определению идиомы // Макет словарной статьи для Автоматизированного толково-идеографического словаря фразеологизмов. М.: Ин-т русского языка, 1991. С. 7-17.

2. Большаков И.А. Многофункциональный словарь-тезаурус для автоматизированной подготовки русских текстов // НТИ сер. 2. 1994. - N1. - С.11 -23.

052-Luckashevich:_Layout 1 18.05.2008 15:49 Page 344

3. Браславский П.И., Соколов Е.А. Автоматическое извлечение терминологии с использованием поисковых машин Интернета // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» (Бекасово, 30 мая - 3 июня 2007 г.). - М.: Изд-во РГГУ, 2007. 658 с.

4. ГОСТ.7.25.2001. Тезаурус информационно-поисковый одноязычный: Правила разработки: структура, состав и форма представления: Межгосударственный стандарт 7.25. - Минск: Межгосударственный совет по стандартизации, метрологии и сертификации, 2001.

5. Добров Б.В., Лукашевич Н.В., Синицын М.Н., Шапкин В.Н., Разработка лингвистической онтологии для автоматического индексирования текстов по естественным наукам // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Седьмой Всероссийской научной конференции (RCDL'2005)– Ярославль: ЯрГУ им.П.Г.Демидова, 2005. – С.70-79.

6. Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции. – 2003, с. 201-210.

7. Добровольский Д.О. Зависит ли синтаксическое поведение идиом от их семантики // Компьютерная лингвистика и интеллектуальные технологии // Труды международной конференции «Диалог 2005». - М.: Изд-во РГГУ, 2005.

8. Лукашевич Н.В. Автоматизированное формирование информационно-поискового тезауруса по общественно-политической жизни России // НТИ. Сер.2. - 1995. - N 3. - С.21-24.

9. ANSI/NISO. Guidelines for the Construction, Format, and Management Monolingual Thesauri. - 2003.

10. Bentivogli L., Pianta E. Extending WordNet with Syntagmatic Information // Proceedings of International Wordnet Conference (GWC - 2004). - 2004. - pp. 47-53.

11. Calzolari N., Fillmore Ch., Grishman R, Ide N., Lenci A., MacLeod C., Zampolli A. Towards Best Practice for Multiword Expressions in Computational Lexicons // Proceedings of LREC - 2002, pp.1934-1940

12. Pearce D. Synonymy in collocation extraction // Proceedings of the NAACL'01 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations – 2001.

13. Sag I., Baldwin T., Bond F., Copestake A., Flickinger D. Multiword expressions: A Pain in the Neck for NLP // Proceedings of CICLING 2002, Mexico city, Mexico. – 2002.

14.Snow R., Jurafsky D., Ng A.Y. Semantic taxonomy induction from heterogenous evidence // Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL. - Sydney, Australia. – 2006. pp.801-808.

053-Ljashevskaja-SharovSA:_Layout 1 18.05.2008 16:23 Page 345

ЧАСТОТНЫЙ СЛОВАРЬ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА: КОНЦЕПЦИЯ И ТЕХНОЛОГИЯ СОЗДАНИЯ

FREQUENCY DICTIONARY OF THE RUSSIAN NATIONAL CORPUS: PRINCIPLES AND TECHNOLOGY

Ляшевская О.Н. *(olesar@mail.ru), Институт русского языка им. В.В. Виноградова РАН*
Шаров С.А. *(s.sharoff@leeds.ac.uk), Университет Лидса, Великобритания*

Словарь содержит представительный базовый словник современного русского языка (2-я половина XX – начало XXI вв.), снабженный информацией о частотности употребления, статистическом распределении по текстам и жанрам, по времени создания текстов. Словарь основан на текстах Национального корпуса русского языка объемом 100 млн. словоупотреблении.

1. Введение

Для русского языка было разработано несколько частотных словарей. Пионером был словарь Г. Йоссельсона, изданный в 1953 году в Детройте на материале языка по преимуществу дореволюционной России. Словари Э.А. Штейнфельд (1963), Л.Н. Засориной (1977), Л. Леннгрена (1993) и др. были созданы на основе относительно небольших коллекций текстов (400 тысяч - 1 миллион слов) и в большой степени отражают специфику русского языка советского периода: частоты слов *товарищ* и *партия* в них сопоставимы со служебными словами, а слово *расческа* отсутствует. Существуют также специализированные словари, в частности, словарь Е.М. Степановой (1976), посвященный общенаучной лексике. Отдельную отрасль статистических словарей составляют словари языка Пушкина, Достоевского, Грибоедова, Цветаевой (Виноградов 1956-1961, Шайкевич и др. 2003, Поляков 1999, Белякова и др. 1996), которые полностью описывают язык данного писателя.

Новый частотный словарь – универсальный. Несмотря на то, что последний его прямой предшественник был выпущен 15 лет назад (Леннгрен 1993), очевидно, что за это время изменилось многое – как сам язык, так и технология подготовки частотных словарей. Наш словарь призван представить статистическую картину современного словоупотребления (1950-2005 г.), заполнив, в частности, лакуну последних двух десятилетий, а также показать изменения, произошедшие в языке с 1950 года.

Словарь базируется на 100-миллионном корпусе, в то время как предыдущие словари опирались на материал объемом от 400 тыс. до 1 млн. словоупотреблений. Национальный корпус (www.ruscorpora.ru, НКРЯ 2005) более представителен по охвату материала, так как содержит сбалансированную коллекцию текстов разных типов, жанров и стилей, в том числе и тексты русского зарубежья. Распределение текстов в подкорпусе современного русского языка (с 1950 года) по функциональным стилям показано в таблице 1. Тексты нехудожественной литературы относятся к более чем 50 предметным областям (экономика и финансы, право, путешествия и др.), а их типология варьируется от законов и научных статей до интервью, инструкций и объявлений (всего более 100 типов). Художественные тексты включают романы, повести, рассказы, очерки, пьесы, сказки, эссе, литературные письма и др.

Художественная литература	36%
Публицистика	42%
Прочая нехудожественная литература	17%
Устная литература	5%

Таблица 1. Функциональные стили подкорпуса современного русского языка

Большой размер и стилистическая сбалансированность корпуса являются предпосылкой того, что он будет давать надежные статистические результаты для наиболее частотных слов: так, состав первых 20 000 элементов не будет существенно меняться, если, сохранив пропорцию, заменить данные тексты другими или сравнить несколько подвыборок корпуса. Это показывает опыт составления частотных словарей других 100-миллионных

053-Ljashevskaja-SharovSA:_Layout 1 18.05.2008 16:23 Page 346

национальных корпусов, таких как британский, чешский (Leech et al. 2001, Čermák & Křen 2004), а также корпуса испанского языка (Davies 2005). Естественно, что частотный словарь НКРЯ во многом, и в технологических вопросах, и содержательно, ориентируется на эти образцы.

2. Размер корпуса и надежность выборки

Существующие частотные словари для русского языка были построены на сравнительно небольших корпусах: ЭВМ первых поколений не могли работать с корпусами большего размера. Интересно, что теоретические рекомендации, выработанные в 1970-е годы (Пиотровский и др. 1972), также доказывали, что для достоверного описания 1600-1700 наиболее частотных слов достаточно использовать корпус размером 400 тыс. словоупотреблений. Эта аргументация строилась на понятии доверительного интервала, который широко используется в статистике и социологии: если мы знаем размер выборки и экспериментальную вероятность события в этой выборке (т.е. частоту слова нашем корпусе), то мы можем вычислить доверительный интервал вероятности этого события на всей популяции (т.е. частоту употребления того же слова во всем пространстве языка).

В таблице 2 приводятся примеры частоты отдельных слов в словарях Леннгрена, Засориной и Штейнфельд в сравнении с частотами НКРЯ и 150-миллионного корпуса русского языка, собранного из Интернета (о последнем см. Sharoff 2006). Несмотря на то, что слова *думать, задача, любить* безусловно относятся к ядру языка (входят в число 200-500 самых частотных лемм), в небольших корпусах даже их частота различается весьма существенно. Частота сравнительно менее частотных слов (*загрязнение, изучение, милый*) варьируется в еще больших пределах. Хотя состав Интернет-корпуса довольно существенно отличается от НКРЯ (большим количеством технических текстов и форумов и меньшим количеством художественной литературы), различия в частоте этих единиц между ними не столь велики.

Лемма	Леннгрен	Засорина	Штейнф.	НКРЯ	Интернет
власть	202	364	138	422	428
думать	609	1094	1058	865	818
загрязнение	69	1	0	9	11
задача	499	421	250	228	292
изучение	193	110	0	63	78
любить	415	632	595	549	650
милый	58	242	135	129	110

Таблица 2. Сравнение частоты отдельных слов (среднее на миллион словоупотреблений).

Как видим, теоретические рекомендации относительно достаточного размера корпуса в данном случае оказываются не слишком достоверными. Причина этого кроется в исходных допущениях на нормальное Гауссово распределение частоты слов, в соответствии с которым каждое слово встречается с одинаковой частотой во всех текстах. Если слово встретилось в тексте один раз, то при нормальном распределении это не влияет на вероятность его употребления там во второй раз. Но в реальности это не так. Каждый текст имеет некоторую собственную тему, слова которой в этом тексте будут употребляться намного чаще среднего. В тексте про хоббитов слово *хоббит* будет употребляться так же часто, как и многие служебные слова, что существенно повысит его частоту в корпусе, который будет включать хотя бы один такой текст[1]. В результате частотный список, построенный на основе корпуса, отражает специфику тех текстов, которые попали в него при его составлении.

Таблица 2 показывает несовершенство частотных словарей, построенных на относительно небольших корпусах, но простое увеличение размера корпуса также не гарантирует стабильности результатов. При интерпретации списков частотного словаря надо помнить, что любой корпус, каким бы большим он ни был, является конечным подмножеством потенциально бесконечного множества текстов на данном языке. Любая другая выборка этого подмножества породит несколько другой список, который будет отличаться в своих менее частотных элементах. Корпус большего размера, отражающий большее количество тем и функциональных стилей (кор-

[1] Кеннет Черч называл эту ситуацию проблемой Норьеги (Church 2000), Адам Килгаррифф - whelk problem, от сравнительно редкого английского слова, обозначающего вид моллюска (Kilgarriff 1997).

053-Ljashevskaja-SharovSA:_Layout 1 18.05.2008 16:23 Page 347

пус типа BNC или НКРЯ), обеспечивает хорошую надежность для наиболее частотных элементов. Тем не менее, дальнейшее увеличение объема текстов в ущерб их разнообразию (см., например, проекты создания Гига-корпусов английского и китайского языков, содержащих более миллиарда словоупотреблений новостных текстов, Cieri & Liberman 2002), может приводить к меньшей надежности частотного списка на таких корпусах за счет сдвига их словаря в сторону новостной лексики.

Поскольку задачей частотного словаря является не просто ранжировать слова по их частоте в отдельном корпусе, но и определить лексическое ядро языка, необходимо отделить слова, часто встречающиеся во многих текстах, от тех, чье лексическое поведение подобно словам *Норьега* или *хоббит*, и которые случайно оказались в той или иной позиции частотного списка. Так в Чешском национальном корпусе используется понятие средней уменьшенной частоты (ARF, Average Reduced Frequency), в котором частота слова взвешивается по расстоянию между отдельными словоупотреблениями (Čermak & Křen 2005). Во многих частотных словарях (Леннгрена, Британского национального корпуса, словаря французской лексики в области бизнеса) используется коэффициент D, введенный А. Жуйаном (Juilland et al. 1970), который принимает во внимание как число документов, в которых встречается слово, так и его относительную частоту в этих документах:

$$D = 100 \times (1 - \frac{\sigma}{\mu\sqrt{n-1}})$$

где μ – средняя частота слова по всему корпусу, σ – среднее квадратичное отклонение этой частоты на отдельных документах, *n* – число документов, в которых встречается это слово.

Значение D у слов, встречающихся в большинстве документов, близко к 100, а у слов, часто встречающихся лишь в небольшом числе документов, близко к 0. Частотный список словаря Леннгрена даже отсортирован по значению произведения этого коэффициента на среднюю частоту слова. В связи с тем, что теоретический статус этого произведения неясен, мы не считали целесообразным сортировать наш словарь по нему. Однако его указание для каждого слова дает возможность оценить, насколько оно специфично для отдельных предметных областей. Например, слова *жуткий, специфический* и *сырье* имеют примерно равную частоту (21 употребление на миллион слов), но при этом коэффициент D у *специфический* - 66, *сырье* - 18, а у *жуткий* - 78, что означает, что последнее слово значимо для большего числа предметных областей и (при прочих равных условиях) имеет большие шансы на место в неспециализированном словаре.

3. Структура словаря

Концепция словаря предполагает издание «бумажной» версии с сопутствующим ей электронным вариантом, представляющим частотный словарь в более полном объеме. Словарная часть содержит следующие разделы:

I. Общая лексика
- алфавитный список лемм
- частотный список лемм
- распределение лемм по функциональным стилям:
  - ▶частотный словарь художественной литературы,
    словарь значимой лексики художественной литературы
  - ▶частотный словарь публицистики,
    словарь значимой газетно-новостной лексики
  - ▶частотный словарь другой нехудожественной литературы,
    словарь значимой лексики
  - ▶частотный словарь живой устной речи,
    словарь значимой лексики живой устной речи
- алфавитный список словоформ

II. Части речи
- частотный список имен существительных
- частотный список глаголов
- частотный список имен прилагательных
- частотный список наречий и предикативов
- частотный список местоимений (местоимения-существительные, прилагательные, наречия, предикативы)
- частотный список лемм служебных частей речи

053-Ljashevskaja-SharovSA:_Layout 1 18.05.2008 16:23 Page 348

III. Вспомогательные таблицы

● данные о частотности частеречных классов и другая статистическая информация

IV. Имена собственные и аббревиатуры

● алфавитный список лемм

В алфавитном списке лемм приводится имя леммы, часть речи, общая частота леммы, число документов, в которых она встретилась и коэффициент вариации D. Общая частота характеризует число употреблений на миллион слов корпуса, или ipm (instances per million words). Это делается для того, чтобы упростить сравнение частоты слова в разных корпусах, которые могут довольно сильно отличаться по своим размерам. Например, если слово власть встречается 55 раз в корпусе размером 400 тыс. слов, 364 раза в миллионном корпусе и 40598 раз в 100-миллионном корпусе современного русского языка и 55673 раза в большом 135-миллионном корпусе НКРЯ, то его частота в ipm составит 137.5, 364.0, 372.06 и 412.39, соответственно. Алфавитный список электронного издания включает 60 000 наиболее частотных лемм.

В списке лемм, упорядоченном по частотности, указываются имя леммы, часть речи, общая частота леммы, число документов, коэффициент D и распределение частотности по десятилетиям. Частотный список включает 20 000 самых частотных лемм.

Частотные словари функциональных стилей составлены на основе подкорпусов художественной литературы, публицистики, другой нехудожественной литературы и устной речи. В список включены 5 000 самых частотных лемм этих подкорпусов. Список наиболее типичных лемм для каждого типа текстов был выделен на основе сравнения частоты лемм в таких текстах и в остальном корпусе. В качестве метрики сравнения был использован критерий отношения правдоподобия (log-likelihood), вычисляемый на основе следующей матрицы:

	Подкорпус	Другие тексты	Весь корпус
Частота	a	b	a+b
Размер	c	d	c+d

На основе этой матрицы значение отношения правдоподобия G2 можно вычислить по следующей формуле (Rayson & Garside 2000):

$$G2 = 2(a \ln(\frac{a}{E1}) + b \ln(\frac{b}{E2})); \text{ где } E1 = c\frac{a}{c}$$

Словари значимой лексики для разных функциональных стилей включают по 500 лемм.

Алфавитный список словоформ включает все словоформы корпуса с частотой выше 0.1 ipm (всего около 15 тыс.); приводится общая частота словоформы. Омонимичные словоформы помечаются знаком *.

В разделе «Части речи» частотный список лемм разбит на шесть подсписков: имена существительные, глаголы, имена прилагательные, наречия и предикативы, местоимения и служебные части речи. Для каждой леммы указана ее общая частота и ранг (порядковый номер) в общем списке. Каждый список содержит по 1 тысяче наиболее частотных лемм.

Вспомогательные таблицы включают в себя данные о частотности частеречных классов, других грамматических категорий, а также информацию о покрытии текста лексемами, средней длине слова, словоформы и предложения.

Завершает словарь алфавитный список имен собственных и аббревиатур. Имена собственные отделены от основной части словника, так как образуют значительно менее стабильную в статистическом отношении группу, а их частотность в большой степени зависит от выбора текстов в корпусе и их хронотопа. В Леннгрен 1993 высказано мнение, что включение имен собственных в частотный словарь на общих основаниях неизбежно приводит к его преждевременному устареванию.

Для получения списка имен собственных и аббревиатур из конкорданса корпуса были выделены имена существительные и сокращения, написание которых в текстах с большой буквы превышало 95-процентный порог, ср. *Россия, Смирнов, ГРЭС, МИД, КЗоТ*.[2] В словарь включена ядерная часть этого списка, насчитывающая 3 000 наиболее частотных единиц.

По традиции, сложившейся для изданий такого рода, на страницах словаря представлена рубрика «Интересные факты»: публикуются списки самых популярных слов различных лексических групп (дни недели, погодные явления, цвета, глаголы движения и т.д.), а также самые длинные словоформы и частотный список знаков пунктуации.

[2] Особо отметим, что прилагательные типа *Христов, Петин, Костромской/костромской* относятся к общей лексике.

053-Ljashevskaja-SharovSA:_Layout 1 18.05.2008 16:23 Page 349

6429	костюм	2288	плащ
4890	сапог	2179	юбка
3696	пальто	1904	шинель
3696	рубашка	1894	наряд*
3410	куртка	1822	туфля
3396	шапка	1668	рубаха
3126	ботинок	1633	джинсы
3041	платок	1585	перчатка
2962	пиджак	1522	шуба
2955	брюки	1356	мундир
2840	штаны	1251	фуражка
2686	шляпа	1235	свитер
2617	берет	1134	валенок

Таблица 3. Частотный список обозначений одежды и обуви.

В качестве примера в таблице 3 мы приводим частоты имен существительных, обозначающих одежду и обувь. Как можно ожидать, список отражает, с одной стороны, «типичность» элементов гардероба (*валенки* занимают только 26 место в списке), а с другой стороны, их «значимость» при описании внешности человека в текстах (*костюм* – более перцептивно выделенная вещь, чем *ботинки*).

4. Подготовка словарного материала

Базовые списки частотного словаря были получены в автоматическом режиме, при этом использовалась метатекстовая и лексико-грамматическая разметка корпуса. На основе метатекстовой информации были построены и сравнивались между собой частотные списки на отдельных выборках корпуса (по функциональным стилям, по времени создания текста). Другой вид разметки, лексико-грамматическая, позволяет установить исходную форму слова (лемму), ее часть речи и такие грамматические характеристики, как падеж, число, время и т. д.[3] Это дало возможность собрать данные о частотности не только отдельных словоформ, но и лексем, а также об употребительности тех или иных грамматических категорий. При создании настоящего словаря был использован вариант лексико-грамматической разметки корпуса с автоматическим разрешением морфологической омонимии.

Русский язык как язык с богатым словоизменением создает дополнительные трудности для составителей частотного словаря, так как многие словоформы в текстах омонимичны (ср. словоформу *стали* как форму глагола *стать* и существительного *сталь*, словоформу *банка*, представляющую леммы *банк* и *банка*, слова типа *вера* и *Вера*). Тем не менее, в частотном словаре исходная форма слова, или лемма, должна быть приписана любой словоформе однозначно.

В словарях предшествующего поколения (Засорина 1977, Леннгрен 1993) омонимия разрешалась вручную, так как объем обрабатываемого корпуса был незначителен. Очевидно, что для 100-миллионного корпуса такое решение не подходит. При составлении настоящего словаря был учтен опыт чешских коллег, которым пришлось дорабатывать морфологический анализатор, пополнять словарь и проводить ручную редактуру. Первоначально корпус НКРЯ был размечен морфологическим анализатором Mystem (Сегалович, Маслов 1998). Неоднозначность в лексико-грамматической разметке была разрешена с помощью программы А.В. Сокирко, использующей модель триграмм и тренировочный подкорпус со снятой вручную омонимией (Сокирко, Толдова 2005).

Существенную проблему для лемматизации представляют также несловарные слова (Ляшевская и др. 2007). Если слово отсутствует в грамматической словаре морфологического парсера, то ему приписываются одна

[3] Принципы лемматизации и состав частей речи определяются морфологическим стандартом корпуса (НКРЯ 2005), который в общем и целом соответствует принципам Грамматического словаря русского языка (Зализняк 1977). Некоторые особенности лемматизации связаны с тем, что сбор данных происходит по преимуществу в автоматическом режиме. Отметим, что учитывается только пословная разметка: устойчивые обороты, составные предлоги и другие неоднословные лексические единицы (ср. *Новый год, в течение, тем не менее, друг друга*) не включаются в словарь.

053-Ljashevskaja-SharovSA:_Layout 1 18.05.2008 16:23 Page 350

или несколько гипотез об исходной форме слова и его грамматических характеристиках. В результате в частотный словарь попадают такие «леммы», как *благодарностий* (ср. словоформу *благодарностию*), *Янсный* (ср. *Янсен*), *Барклаивать* (ср. *Барклай*). Между тем, доля несловарных словоформ в НКРЯ составляет 3% всех словоупотреблений и 45% списка словоформ корпуса. Для частотных несловарных словоформ использовались программы пост-обработки морфологической разметки НКРЯ, составленные Б.П. Кобрицовым и Г.К. Бронниковым, а также результаты валидации работы этих программ, полученные О.Н. Ляшевской и Д.К. Бронниковой (Ляшевская 2007, Бронникова 2007). Наиболее эффективными оказались два подхода к лемматизации несловарных слов: кластеризация гипотез о лемме и типе парадигмы (наиболее вероятным для словоформы считается тот разбор, который встречается и у других несловарных словоформ, таким образом, словоформы «ищут» себе соседей по словоизменительной парадигме) и выделение наиболее продуктивных приставок.

Поскольку автоматическое разрешение омонимии и интерпретация несловарных форм допускают определенную, хотя и незначительную, погрешность, омонимы, входящие в первые 20 тысяч частотных слов, подверглись дополнительной ручной проверке.

***

Авторы выражают благодарность В.А. Плунгяну, А.Я. Шайкевичу, а также Е.А. Гришиной, Б.П. Кобрицову, Е.В. Рахилиной, Д.В. Сичинаве и другим участникам семинара НКРЯ, принимавшим участие в обсуждении принципов создания словаря. Мы благодарим О. Урюпину, Д. и Г. Бронниковых, Б. Кобрицова, сотрудников ООО «Яндекс» А. Аброскина, Н. Григорьева, А. Сокирко за помощь в сборе и обработке материала.

Список литературы

1. Бронникова Д.К. Сравнение алгоритмов лемматизации на материале Национального корпуса русского языка. Дипломная работа. М.: РГГУ, 2007.

2. Белякова И.Ю., Оловянникова И.П., Ревзина О.Г. (сост.). Словарь поэтического языка Марины Цветаевой. В 4-х томах. М: Дом-музей Марины Цветаевой, 1996.

3. Виноградов В.В. (отв. ред.). Словарь языка Пушкина. Т. I – IV. М., 1956-1961.

4. Зализняк А.А. Грамматический словарь русского языка: Словоизменение. М., 1977; 4-е изд.: М.: Русские словари, 2003.

5. Засорина Л.Н. (ред.). Частотный словарь русского языка. Москва: Русский язык, 1977.

6. Лённгрен Л. (ред.). Частотный словарь современного русского языка [Lönngren, Lennart. The Frequency Dictionary of Modern Russian. Acta Univ. Ups., Studia Slavica Upsaliensia Uppsala 32]. Uppsala, 1993.

7. Ляшевская О.Н.. К проблеме лемматизации несловарных слов // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007». М, 2007.

8. Ляшевская О.Н., Кобрицов Б.П., Сичинава Д.В. Автоматизация построения словаря на материале массива несловарных словоформ // Интернет-математика 2007. Екатеринбург, 2007.

9. НКРЯ: Национальный корпус русского языка 2003-2005: Результаты и перспективы. М.: Индрик, 2005.

10. Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.А.. Математическая лингвистика. М.: Высшая школа, 1972.

11. Поляков А.Е.. Электронный словарь языка писателя (на примере языка А.С. Грибоедова) // Труды Международного семинара Диалог-99 по компьютерной лигвистике и ее приложениям. Таруса, 1999. М., 1999. Т. 2. С. 230-236.

12. Сегалович И., Маслов М.. Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов // Труды международной семинара Диалог'98 по компьютерной лингвистике и ее приложениям. Казань, 1998. Т.2. С. 547–552.

13. Сокирко А.В., Толдова С.Ю. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка // Международная конференция «Корпусная лингвистика 2004». С.-Пб., 2004.

14. Степанова Е.М. Частотный словарь общенаучной лексики. М., 1976.

15. Шайкевич А.Я., Андрющенко В.М., Ребецкая Н.А. Статистический словарь языка Достоевского. М.: Языки славянской культуры, 2003.

16. Штейнфельд Э.А. Частотный словарь современного русского литературного языка. Таллин, 1963.

17. Čermák F., Křen M. (eds.). Frekvenční slovník češtiny (Frequency dictionary of Czech). Praha: NLN, 2004.

18. Čermák F., Křen M. New generation corpus-based frequency dictionaries: The case of Czech // International Journal of Corpus Linguistics, 10, 2005. P. 453-467.

19. Church K.W. Empirical estimates of adaptation: the chance of two Noriegas is closer to p/2 than p^2 // Proceedings of the 18th Conference on Computational Linguistics (COLING). Saarbrücken, Germany, 2000. Vol. 1. P. 180-186.

053-Ljashevskaja-SharovSA:_Layout 1 18.05.2008 16:23 Page 351

20. Cieri Ch., Liberman M. Language resources creation and distribution at the Linguistic Data Consortium // Proceedings of LREC 02. Las Palmas, Spain, 2002. C. 1327-1333.

21. Davies M. A Frequency Dictionary of Spanish: Core Vocabulary for Learners. London – N.Y.: Routledge, 2005.

22. Josselson H.H. The Russian Word Count and Frequency Analysis of Grammatical Categories of Standard Literary Russian. Detroit: Wayne University Press, 1953.

23. Juilland A., Brodin D., Davidovitch C. Frequency Dictionary of French Words. The Hague-Paris: Mouton, 1970.

24. Kilgarriff A. Putting frequencies in the dictionary // International Journal of Lexicography, 10 (2), 1997. P. 135-155.

25. Leech G., Rayson P., Wilson A. Word Frequencies in Written and Spoken English: based on the British National Corpus. London: Longman, 2001.

26. Rayson P., Garside R. Comparing corpora using frequency profiling // Proceedings of the Comparing Corpora Workshop at ACL 2000. Hong Kong, 2000. P. 1-6.

27. Sharoff S. Creating general-purpose corpora using automated search engine queries // Baroni M., Bernardini S. (eds.), WaCky! Working papers on the Web as Corpus. Bologna: Gedit, 2006. http://wackybook.sslmit.unibo.it.

054-Marcasova_EV:_Layout 1 18.05.2008 15:50 Page 352

РИТОРИЧЕСКАЯ ЭНАНТИОСЕМИЯ В КОРПУСЕ РУССКОГО ЯЗЫКА ПОВСЕДНЕВНОГО ОБЩЕНИЯ «ОДИН РЕЧЕВОЙ ДЕНЬ»

RHETORICAL ENANTIOSEMY IN THE SPEECH CORPUS OF THE RUSSIAN EVERYDAY COMMUNICATION “ONE SPEAKER’S DAY”

Маркасова Е.В. *(markasovaelena@yandex.ru)*
Санкт-Петербургский государственный университет

В докладе анализируется энантиосемия, опознаваемая лишь в устной речи. В отличие от ингерентной и адгерентной энантиосемии, выделяемый нами вид характеризуется не внутренней антонимией значений, а противопоставленностью коммуникативных установок (конструктивной и деструктивной).

Ключевые слова: *энантиосемия, спонтанная речь, речевая агрессия, интонация*.

Введение

Энантиосемия, совмещение в слове противоположных значений, «внутренняя антонимия», оценивается учеными по-разному: одни считают его непродуктивным явлением, «реликтом семантики древних корней» [Новиков 1973: 192], другие называют регулярным языковым явлением [Шмелев 2000, Ермакова 2002]. Исследователи современного русского языка разграничивают ингерентную энантиосемию, которая реализуется в совмещении значений в слове и находит отражение в словарях, и адгерентную энантиосемию, которая не отражается в словарях, демонстрирует изменение коннотации и, как правило, сопровождается иронией [Цоллер 1998]. Примеры ингерентной энантиосемии: *законник* - «вор в законе» и *законник* - «человек, не отступающий от буквы закона», *прослушать* - «внимательно выслушать» и *прослушать* - пропустить мимо ушей», наверное - «может быть» и наверное - «совершенно точно», *вывести* - «уничтожить» (вывести тараканов) и *вывести* - «создать» (вывести сорт). Описание полной и частичной энантиосемии (актантной, коннотативной, основанной на разнонаправленности прагматических компонентов) см. в: [Ермакова 2002].

В исследованиях энантиосемии ученые традиционно используют термины риторики *ирония* и *антифразис* [Балли 1955; Клюев 1999; Москвин 2007], представляющиеся удобными для комментирования тех случаев энантиосемии, при которых положительное коннотативное значение заменяется отрицательным:

(1) *Опять двойка? Ты* ***супер****!* ***Какая радость****!*

Использование бранной лексики с целью характеристики чего-либо исключительного, вызывающего восхищение, называется в риторике *астеизмом.* Это явление, противоположное *антифразису,* также распространено и в разговорной речи, и в художественных текстах. (Отметим, что не все исследователи считают подобные примеры энантиосемией. См. [Ермакова 2002]). Например:

(2) *И этот* ***извращенец*** *ходит в библиотеку к девяти утра и сутками пашет!*

Адгерентная энантиосемия всегда жестко связана с внутренней иерархией участников коммуникативного акта [Скляревская 1994; Цоллер 1998], причем как антифразис, так и астеизм предполагают, по меньшей мере, наличие долгого опыта общения адресата и адресанта, знание обоими ситуации и наличие общих этических ориентиров. Иными словами, если адресат примера (2) считает, что любой интеллектуальный труд - бессмысленная трата времени, то он поймет эту реплику как содержащую отрицательную оценку описываемого. Обратим внимание и на тот факт, что такой вид энантиосемии часто сопровождается выветриванием значения слова, оно становится чистым выразителем эмоциональной оценки, подобно междометию. Астеизм, как и антифразис, может служить формой проявления речевой агрессии [Ермакова 1997; Шаронов 2004].

Названные случаи энантиосемии описаны лингвистами. Однако рассмотрение этого явления на материале звукового корпуса русского языка повседневного общения «Один речевой день» привело нас к мысли о том, что есть еще один тип энантиосемии, часто используемый в разговорной речи, но не получивший освещения в научной литературе. Отличительные особенности этого типа характеризуются в нашей работе.

054-Marcasova_EV:_Layout 1 18.05.2008 15:50 Page 353

Материал

В корпусе «Один речевой день», созданном и пополняемом группой исследователей в Санкт-Петербургском государственном университете, при расшифровке и описании материалов, поступивших в ноябре 2007 г., были выделены записи для прослушивания.

Совершенно очевидно (об этом свидетельствуют и наблюдения над лексикой данных фрагментов), что И19 (женщина 30-35 лет, в общей системе обозначения информантов в базе «Один речевой день» именуемая И19) не говорит ничего обидного, желает собеседнику удачи, даже хвалит (*молодец, замечательно, хорошо, отлично)*. Однако при этом имя ребенка и другие маркеры доверительности (типа *солнышко, заинька, зайка, котик, умница, дорогой, милый, миленький и пр.)* произносятся без характерных для разговора с детьми и свойственных доброжелательному адресанту особенностей: нет ни продления долготы звука сверх обычного для ударных, ни варьирования частоты основного тона, ни повышения регистра, ни гиперлабиализованности [Гаврилова 2001; Кодзасов, Кривнова 1977], ощущается отсутствие эмоциональной составляющей.

В следующих далее фрагментах текста адресант (использует обращения, традиционные для разговоров с близкими людьми (*солнышко, (мое) солнце, (моя) радость),* а также само имя ребенка с уменьшительно-ласкательными суффиксами *оньк, очк*:

(3) *Але-о!/ Привет / что все?/ закончилось / у вас закончилось /что случилось?/ почему? / ты где? то есть она сейчас уже ушла? а спроси у Иры /да-а/ я же тебе сказала /подойти к Ирине и спросить/ попросить помощи / Да / дойди / дойди* ***солнышко****/ давай/ удачи/ ну тебе/ у тебя все хорошо/ ну замечательно/ ну /угу /хорошо /****солнышко*** */ давай /не теряй времени /подойти к Ирине /попроси помощи найти Нину Филипповну/и отзвонись мне/ пока //*

(4) *Але /да* ***солнышко*** */ну так /ну замечательно/ отлично /поздравляю тебя /* ***молодец*** */все не так страшно /готовься /что делать/ Держись /держись мое* ***солнце*** */ну давай / держись / пока //*

(5) *Ты выпила* ***водичку****? Отлично! <…> Что случилось? Что именно ты забыла, моя* ***радость****? Можно поподробнее? Что ты забыла? Что сегодня у вас был английский/ ты забыла. <…>Я не поняла, что да? Зачет?<…>проверка по словам<…> Какие слова? Что/ и сейчас забыла? И сейчас забыла / какие слова? <…> Из какой лексики/* ***Лизонька****?*

(6) ***Лизонька*** */ это Марина Викторовна ваша / такое дурацкое слово употребляет «лексика» /оно дурацкое /* ***Лизочка*** */ Лексика это всего-навсего слова <…>Лексику вы учите / бред какой /*

(7) ***Моя девочка/*** *<…>ну что делать-то с этой двойкой /****моя хорошая***

Мы предположили, что особенности интонирования маркеров доверительности создают конфликт горизонта ожиданий слушающего и интенций говорящего, а диссонанс между семантикой слова и нейтральной интонацией (при ожидаемой экспрессивной) становится основой для аномального эмоционального фона при общении. Для того чтобы проверить эту гипотезу, был проведен эксперимент.

Эксперимент

Участникам эксперимента (группа из 30 студентов и школьников) было предложено прочитать расшифровку и ответить на такие вопросы: «Часто ли родители так разговаривают с детьми? Типично ли содержание разговора? Что можно сказать об этой женщине?» Информанты восприняли текст как банальный, многие предположили, что ребенок должен сдавать какой-то экзамен или зачет, которого боится, что в связи с этим мама очень переживает, волнуется, старается поддержать дочку.

После прослушивания записи предлагалось ответить на вопрос: «Что нового вы узнали о Лизе?» Информанты реагировали крайне эмоционально, причем не отвечали на поставленный вопрос, а выражали мнение по поводу высказываний И19: «Почему она таким прокурорским тоном разговаривает?» «А Вы можете на нее повлиять, чтобы она так больше не говорила?» «Вот пойдет в школу у Вас сын, Вы, может, тоже еще так заговорите!» «Эта женщина, наверное, просто устала, а дочка у нее еще неизвестно что такое.» Реплики, при всем разнообразии оценок поведения Лизы и ее матери, отражают одно: в прослушанных фрагментах участники эксперимента ощутили нечто неприятное, раздражающее, чего невозможно уловить при письменной передаче текстов.

Это явление может быть описано как разновидность внутрисловной антонимии, риторическая энантиосемия.

Риторическая энантиосемия

Термином «риторическая энантиосемия» мы предлагаем обозначать случаи конфликтного совмещения в слове (словосочетании, предложении) разнонаправленных коммуникативных установок

054-Marcasova_EV:_Layout 1 18.05.2008 15:50 Page 354

(контактоустанавливающей и деструктивно-агрессивной), при котором семантика рассматриваемой единицы (включая оценочную составляющую) не претерпевает никаких изменений: *Лизонька* остается *Лизонькой*. Значения слов с риторической энантиосемией, участвующих в коммуникативных актах, не претерпевают никаких трансформаций: не происходит ни мелиорации, ни пейоративизации значений, не актуализируются непрямые значения. Вместе с тем, здесь нет и интонационного отрицания называемых признаков или фактов.

Даже лишенная интонации угрозы или иронии, фраза, включающая риторическую энантиосемию, создает напряженный эмоциональный фон. Не случайно при прослушивании звукового файла продолжительностью 22 минуты 36 секунд (продолжительность разговора с матерью) девочка пытается заплакать семь раз. Многочисленные повторы ласковых слов (примеры 5, 6, 7), произносимых нейтральным тоном, способствуют нарастанию напряжения, а затем разрешаются каскадом вопросов или жалобами.

Примеры (3) и (4), фрагменты разговора по телефону, не дают возможности проследить за реакцией адресата, однако в них отсутствие эмоциональной окрашенности также очевидно.

Аналогичные случаи интонационного оформления обращений, воспринимаемых как агрессивные, мы записали в школе и детском саду:

(8) *Наша* ***деточка*** *начнет учиться / и причесываться /правда ?*

(9) *Все /* ***дорогие мои*** */ быстро одеваться и строиться!*

Мы считаем, что именно «нейтральность» интонирования анализируемых лексем в записях И19 делает их способом проявления речевой (вербальной, коммуникативной) агрессии, такого типа конфликтного речевого поведения, при котором целью агрессора является подавление собеседника через снижение его самооценки.

Одним из способов выражения речевой агрессии является использование маркеров чуждости, к которым относятся, например, характеристики *всякий, разный, какой-то, кто-то, не пойми какой, бог знает какой и др.,* содержащие оценку «несущественный, не заслуживающий внимания». Особую группу маркеров чуждости составляют слова и выражения, показывающие недоверие к адресату: *якобы, будто бы, как бы, так называемые, с позволения сказать, как Вы выразились.* [Шейгал 2004; Пеньковский 1989]

Именно к маркерам чуждости и следует отнести ту особую интонацию отстраненности, отчужденности, которая характеризует примеры с (3) по (7).Отметим, что в записях не наблюдается таких проявлений агрессивности речевого поведения, как сверхполный тип произношения, понижение тона, повышение голоса с целью оказать давление на собеседника [Крейдлин 2000: 483], эксплуатация ИК-7 с целью передать отрицание [Брызгунова 1980: 118-120; Светозарова 1982], а также конструкций, напоминающих угрозатив (термин А.Летучего) [Летучий 2007] .

В отличие от нормального (не агрессивного) речевого поведения, при котором собеседники чувствуют себя равноправными, в условиях речевой агрессии исключается равноправие участников диалога, один из них становится агрессором, другой – жертвой. Видимо, в репликах И19 и проявляется агрессия ради агрессии, с помощью которой за счет близких людей снижается эмоциональное напряжение говорящего.

В основном мы говорили о случаях, когда слова с положительной коннотацией, требующие эмоционально окрашенной интонации, произносятся нейтральным тоном.

Вполне резонно предположить, что есть и противоположный вариант: слова с отрицательной коннотацией, произносимые с особой «положительно окрашенной» интонацией, теряют свое пейоративное значение. Вот, например, обращение хозяйки к кошке с котятами:

(10) *Ах ты краса-авица* ***/ сво-олочь ты /со сволоченя-атами*** *//*

Этот пример можно, используя термины риторики, назвать интонационным астеизмом, причем, в отличие от традиционного астеизма, который уничтожает лексическую составляющую и возникает лишь на основе семы интенсивности, в примере (10) сохраняется лексическое значение: кошка родила котят, когда хозяйка должна была уезжать в отпуск (поэтому и «сволочь», разрушившая все планы).

В свою очередь, отсутствие нейтрального (в норме) тона, замена его доверительно-дружеским, в определенных ситуациях воспринимаются как угроза. Например:

(11) *А он* (начальник - Е.М) *мне говорит / такой / улыбается /* ***Наташенька*** */ давайте-ка/ кофейку / я думаю / чо же это он такой ласковый / ага /добрый // надо выйти в субботу//*

Рассмотренные примеры не могут быть названы ни антифразисом (вследствие отсутствия иронии), ни астеизмом (из-за отсутствия в наших примерах в качестве доминирующих не только состояний восхищения или восторга говорящего, но и фоновых положительных эмоций).

Перспективы

Итак, мы предлагаем называть риторической энантиосемией случаи совмещения в слове (словосочетании, предложении) контактоустанавливающей и деструктивно-агрессивной коммуникативных установок, при

054-Marcasova_EV:_Layout 1 18.05.2008 15:50 Page 355

котором семантика рассматриваемой единицы (включая оценочную составляющую) не меняется.

Риторическую энантиосемию необходимо рассматривать отдельно от тех примеров, которые демонстрируют различные возможности совмещения или трансформации значений (мелиорации, пейоративизации, актуализации непрямых значений и др.).

Продолжение работы с материалом звукового корпуса «Один речевой день» даст возможность прояснить следующие вопросы:

• существует ли возможность создать формальное описание риторической энантиосемии;

• какие части речи и члены предложения наиболее склонны к проявлению риторической энантиосемии;

• действует ли применительно к этому типу энантиосемии установленная В.Н. Цоллером закономерность («в русском языке более частотна пейоративация (смена знака оценки от «+» к «– »), тогда как «случаи реверсии аксиологического знака от минуса к плюсу менее распространены»);

• чем отличается нейтральность интонации воспитанного человека от агрессивной нейтральности интонации;

• как связана способность к «исполнению» высказываний, содержащих примеры риторической энантиосемии, с психологическими особенностями и социальным статусом адресата и адресанта.

Список литературы

1. Балли Ш. Общая лингвистика и вопросы французского языка. М.,1955.

2. Брызгунова Е.А. Интонация // Русская грамматика. Т. I. М.: Изд-во АН СССР, 1980.

3. Гаврилова Т.О. Регистр общения с детьми (baby-talk): некоторые особенности интонации // Антропология. Фольклористика. Лингвистика: Сб. статей. СПб., Изд-во ЕУ в СПб. 2001. С. 227-238.

4. Ермакова О.П. Об иронии и метафоре // Облик слова: Сб. статей памяти Дмитрия Николаевича Шмелева. М.: АО «Астра семь», 1997. С. 48-58.

5. Ермакова О.П.Существует ли в русском языке энантиосемия как регулярное явление? Вспоминая общую этимологию начала и конца // Логический анализ языка. М.: «Индрик», 2002. С. 61-68.

6. Клюев Е.В. Риторика. Инвенция. Диспозиция. Элокуция. М.: ПРИОР, 1999.

7. Кодзасов С.В. Фонетика Интенсификации 2001 // http://www.philol.msu.ru/~otipl/SpeechGroup/publications/kodzasov/intens.rtf.

8. Кодзасов С.В., Кривнова О.Ф. Фонетические возможности гортани и их использование в русской речи / Проблемы теоретической и экспериментальной лингвистики. Изд-во МГУ. 1977.

9. Летучий А.Б. Русский «угрозатив» и его родственники // http://www.dialog-21.ru/dialog2007/materials/html/57.htm

10. Москвин В.П. Выразительные средства современной русской речи: тропы и фигуры. Ростов-на-Дону: Феникс, 2007. С.128-129.

11. Новиков Л. А. Антонимия в русском языке (Семантический анализ противоположности в лексике). М.: Высшая школа, 1973.

12. Пеньковский А.Б. О семантической категории «чуждости» в русском языке //Структурная лингвистика. 1985-1987. М.: Наука, 1989. С. 54-82.

13. Светозарова Н.Д. Интонационная система русского языка. Л.: Изд-во ЛГУ, 1982.

14. Скляревская Г.Н. Новый академический словарь. Проспект. СПб.: ИЛИ РАН, 1994. С. 51 – 52.

15. Цоллер В.Н. Эмоционально-оценочная энантиосемия в русском языке // Филологические науки. М., 1998. № 4. С. 79-80.

16. Шаронов И.А. Приемы речевой агрессии: насмешка и ирония //Агрессия в языке и речи: Сб. статей под ред. И.А.Шаронова. М.: РГГУ, 2004. С.38-52.

17. Шейгал Е. Семиотика политического дискурса. М.: ИТДК «Гнозис», 2004.

055-Mitrenina_OV:_Layout 1 18.05.2008 15:51 Page 356

СИНТАКСИС КОРРЕЛЯТИВНЫХ КОНСТРУКЦИЙ РУССКОГО ЯЗЫКА С ПОЗИЦИИ ГЕНЕРАТИВНОЙ ГРАММАТИКИ

SYNTAX OF CORRELATIVE CONSTRUCTIONS IN RUSSIAN: A GENERATIVE APPROACH

Митренина О.В. *(mitrenina@gmail.com)*
Санкт-Петербургский государственный университет

В работе показано, что в коррелятивных предложениях русского языка наблюдается барьер между коррелятивной и главной клаузой и невозможна реконструкция в главной клаузе. Предлагается предварительная структура коррелятивных предложений русского языка, которая учитывает позиции топика и/или фокуса.

К коррелятивным конструкциям (коррелятивам) принято относить предложения следующего типа:

(1) Кто пришел первым, тот занял лучшие места.

Коррелятивы — это сложноподчиненные предложения с препозицией придаточной части, которая связана со своей вершиной в главной клаузе двойной связью: подчинительной (определительной) и анафорической.

Такое понимание коррелятива сложилось в лингвистике в середине 1990-х годов. Ранее встречалось и более широкое понимание коррелятивных предложений. Так, Эдвард Кинан считал коррелятивами любые относительные конструкции, которые не являлись синтаксически зависимыми от именной вершины [Keenan 1985: 164]. Таким образом, по классификации Кинана к коррелятивам относились и такие предложения, которые не являются коррелятивными с точки зрения современной лингвистики:

(2)	ʔi	čhuya-ø	futma	cephi	šoy̆ikhiʔ	(язык ваппо)
	мне	дом-DO	купил	он(SUBJ)	сгорел’	

‘Дом, который я купил, сгорел’ (пример из [Keenan 1985: 165])

Исторически к коррелятивам иногда относили и конструкции с постпозицией придаточной части, но в современной терминологии такие конструкции более не считаются коррелятивными.

За последние 15 лет в генеративной лингвистике появилась целая серия работ, посвященных формальному анализу синтаксиса коррелятивных конструкций [Izvorski 1996, Vries 2002, Bhatt 2003, Lipták 2005, Dikken 2005].

Раджеш Бхатт приводит список из 25 языков, использующих коррелятивы [Bhatt 2003: 6]. В этом списке встречается древнерусский язык [Keenan 1985], а также южнославянские языки: болгарский, македонский и сербо-хорватский [Izvorski 1996]. Р. Бхатт подчеркивает, что его список не является исчерпывающим. Однако отсутствие в этом списке современного русского языка не является случайным. В современных исследованиях коррелятивные структуры русского языка практически не рассматриваются. Лишь в некоторых работах русскоязычные примеры используются как иллюстрации для тех или иных теорий или классификаций.

Однако, русский язык очень богат коррелятивными конструкциями. При этом коррелятивы встречаются не только в просторечии или в разговорной речи. Анализ корпусных примеров (Национальный корпус русского языка) показывает, что большее число коррелятивов содержится в устной форме литературного языка. Но немало примеров можно найти и в публицистике, и в классических художественных текстах русского литературного языка:

(3) Кто испытал наслаждение творчества, для того уже все другие наслаждения не существуют (*Чехов*) [Шведова 1980: 533].

В современных работах не существует унифицированной системы терминов для обозначения основных элементов коррелятивной конструкции. Коррелятивное предложение можно разбить на два элементарных предложения (клаузы): придаточное предложение и следующее за ним главное предложение. Мы будем называть их, соответственно, **коррелятивной клаузой** (*кто пришел первым* в примере (1)) и **главной клаузой** (*тот занял лучшие места*). В главной клаузе можно выделить **определяемую группу** (*тот*), которая выступает синтаксической вершиной по отношению к коррелятивной клаузе. Определяемая группа может быть **простой** (состоящей из одной словоформы) или **сложной** (состоящей из указательного элемента и именной группы ИГ). Пример коррелятивного предложения со сложной определяемой группой:

(4) Какую хотел, [**такую машину**] и купил.

055-Mitrenina_OV:_Layout 1 18.05.2008 15:51 Page 357

В коррелятивной клаузе можно выделить **относительную группу** (*кто*), которая, как правило, отсылает к тому же объекту, что и определяемая группа. Относительная группа также может быть **простой** (состоящей из одной словоформы) или **сложной** (состоящей из вопросительного элемента и ИГ). Пример коррелятивного предложения со сложной относительной группой:

(5) [**Какую машину**] хотел, ту и купил.

Р. Бхатт вводит также понятие множественного коррелятива, в котором коррелятивная клауза содержит несколько относительных групп, соответствующих такому же числу определяемых групп в главной части. Множественные коррелятивы встречаются и в русском языке:

(6) **Кому какая** премудрость далась, **тот той** и придерживайся! (Тургенев)

Анафорическая связь в коррелятивной конструкции возникает за счет того, что содержание относительной группы воспроизводится в определяемой группе главной клаузы. Относительная и определяемая группы обычно отсылают к одному и тому же объекту, поэтому их принято помечать одинаковыми индексами. К этому же объекту, как правило, отсылает и вся коррелятивная клауза в целом, поэтому и в таком случае ее можно помечать тем же индексом:

(7) [Кто$_i$ пришел первым]$_i$, тот$_i$ занял лучшие места.

В классификации «Русской грамматики» коррелятивные конструкции относятся к сложноподчиненным предложениям нерасчлененной структуры с неориентированной местоименно-относительной (анафорической) связью частей, в которых придаточное предложение предшествует главному [Шведова 1980: 532].

Попытка формального анализа коррелятивов русского языка была предпринята в работе А.А. Зализняка и Е.В. Падучевой [Зализняк, Падучева 1975]. На материале различных языков авторы предложили типологию относительных предложений, в которой свое место заняли и русские коррелятивные конструкции. В этой работе А.А. Зализняк и Е.В. Падучева рассмотрели те свойства относительных предложений, о которых могла идти речь в 1975, с учетом российской традиции синтаксических исследований.

Коррелятивы принято противопоставлять таким определительным предложениям, в которых придаточная часть следует за определяемым элементом. Эти конструкции отличаются от коррелятивов не только порядком следования придаточной и главной частей. Они обладают разными свойствами и не могут быть связаны трансформационно [Bhatt 2003, Lipták 2005].

Эти отличия коррелятивов от стандартных определительных придаточных предложений характерны и для русского языка:

1) В отличие от обычных относительных предложений коррелятивы допускают ИГ как в определяемой, так и в относительной группе, а также в определяемой и относительной группах одновременно:

(8a) Какую **машину** хотел, такую и купил. (ИГ в определяемой группе)

(8b) Какую хотел, такую **машину** и купил. (ИГ в относительной группе)

(8c) Какую **машину** хотел, такую **машину** и купил. (ИГ в относительной и в определяемой группах одновременно).

При этом определительные предложения с предшествующей главной клаузой допускают ИГ только в составе определяемой группы, а в составе относительной группы не допускается никаких именных элементов:

(8d) Купил такую **машину**, какую (***машину/*Тойоту)** хотел.

2) Необходимость указательного элемента:

(9a) Какую ему дают, **ту** пищу он и ест. (коррелятив, указательный элемент и ИГ)

(9b) * Какую ему дают, пищу он ест. (коррелятив, ИГ без указательного элемента)

(9c) Он есть (ту) пищу, которую ему дают. (обычное относительное предложение, возможно наличие или отсутствие указательного элемента)

3) Наличие нескольких относительных и определительных групп допускается только в коррелятивах, но не в других относительных предложениях:

Кто что *хочет,* ***тот то*** *и делает.*

Для разных языков подходящими оказывались разные варианты описания коррелятивных структур. Анализ, предполагающий порождение коррелятивной клаузы внутри главной клаузы с последующим вынесением коррелятивной клаузы в начало предложения, был разработан Р. Бхаттом для простых коррелятивов в хинди. Схема для этого анализа выглядит следующим образом:

(10a) [$_{IP}$ [$_{CorCP}$ ... Rel-XP ...] $_i$ [$_{IP}$ ... (Dem-XP$_j$) [t$_i$] Dem-XP$_j$ ...]]

(10b) [$_{CorCP}$ jo laRkii khaRii hai] *vo* lambii hai

REL девочка стоящая есть та высокая есть

‘Девочка, которая стоит, высокая.’

При таком подходе коррелятивная клауза в глубинной структуре предложения присоединена к определяемому слову в главной клаузе. При переходе к поверхностной структуре она перемещается в начало

055-Mitrenina_OV:_Layout 1 18.05.2008 15:51 Page 358

предложения. Одновременно может происходить факультативное перемещение определяемой группы DemXP.

В пользу передвижения в хинди коррелятивной клаузы из главной клаузы предложения говорят, в частности, такие факты как **возможность реконструкции** внутри главной клаузы и **наличие барьера** между коррелятивной и главной клаузами.

Операция реконструкции применяется при переходе от S-структуры (аналог поверхностной структуры в поздних версиях генеративной грамматики) к логической форме LF (интерфейс, взаимодействующий с концептуально-интенциональной системой реализации языка). Эта операция «возвращает» в исходную позицию элементы, перемещенные вместе с вопросительным или относительным словом. Благодаря реконструкции можно объяснить грамматическую правильность предложений, нарушающих принципы связывания:

(11a) [Какие **свои$_i$** рисунки]$_k$ **Миша$_i$** дал на выставку t_k?

(11b) *[Какие **Мишины$_i$** рисунки]$_j$ **он$_i$** дал на выставку t_j?

В предложении (11a) Анафор *свои* связан именной группой *Миша* благодаря реконструкции, которая позволяет интерпретировать всю группу [Какие свои$_i$ рисунки] в ее исходной позиции в конце предложения. В примере (11b) референциальное выражение *Мишины* оказывается связанным прономиналом *он также* благодаря реконструкции, поскольку группа [Какие Мишины рисунки] интерпретируется в своей исходной позиции в конце предложения, что объясняет грамматическую неправильность этого предложения.

Реконструкция возможна и при скрэмблинге, т.е. таком «факультативном» стилистическом перемещении лексических групп, которое встречающиеся только в языках со свободным порядком слов. При этом перемещенные при скрэмблинге группы интерпретируются в своей исходной позиции:

(12a) [**Свои$_i$** рисунки]$_j$ Миша$_i$ дал на выставку t_j.

(12b) *[**Мишины$_i$** рисунки]$_j$ он$_i$ дал на выставку t_j.

В примерах (12а-б) группы [*Свои рисунки*] и [*Мишины рисунки*] благодаря реконструкции анализируются в своей исходной позиции — в конце предложения. Поэтому в случае (12а) возвратное местоимение *свои* оказывается связанным референтнциальным выражением *Миша*, в результате чего предложение является грамматически правильным. Аналогичным образом референциальное выражение *Мишины* оказывается связанным в (12b), что нарушает принцип С теории связывания, поэтому предложение получается грамматически неправильным.

Р. Бхатт показал, что в коррелятивных предложениях в хинди действует реконструкция [Bhatt 2003]. Подлежащее главной клаузы не может быть кореферентно элементам коррелятивной части:

(13) *[$_{CorCP}$ jo laṛkii Sita-ko$_j$ pyaar kar-tii hai]$_i$ us-ne$_{k/*j}$ os-ko$_i$ ṭhukraa di-yaa

[$_{CorCP}$Которая девушка Ситу$_j$-Асс любит]$_i$ тот$_{k/*j}$ ту$_i$ отверг.

‘Сита отверг ту девушку, которая его любит’. (Более точный перевод: ‘Какая девушка Ситу любит, ту он отверг’).

Эти данные говорят в пользу анализа (10а), который предполагает порождение коррелятивной клаузы внутри главной клаузы.

Р. Бхатт показал **наличие барьера** между коррелятивной и главной клаузами [Bhatt 2003], что также свидетельствует о передвижении элемента при образовании «поверхностного» предложения. Барьером в генеративной грамматике принято называть позицию в предложении, непроницаемую для передвижения элементов.

Так, предложение (14а) является грамматически правильным, а (14b) — грамматически неправильным.

(14a) Who did Mary say [that Bill saw]?

‘(Про) кого Мэри сказала, что Билл (его) видел?’

(14b) *Who did Mary leave [after Bill saw]?

‘Мэри ушла, после того как Билл увидел кого?’

(Букв: ‘Кого Мэри ушла, после того, как Билл (его) увидел?’)

Неправильность предложения (14b) вызвана наличием барьера: вопросительное слово *who* ‘кто’ может пересечь границу придаточного предложения, если оно является дополнением (сентенциальным актантом) глагола в главном предложении, но не может пересечь этой границы, если придаточное предложение — обстоятельство (сентенциальный сирконстант) [Тестелец 2001: 192] .

Аналогичным образом в языке хинди в предложении (15а) передвижение возможно, а в (15b) возникает барьер, и выдвижение элемента из именной группы невозможно:

(15a) [jo laṛki: TV-par ga: rah-i: hai]$_i$ [Sita soch-ti:hai [$_{CP}$ki voi sundar hai]]

[Которая девушка по-ТВ поет]$_i$ Сита думает [$_{CP}$что та красивая есть]]

‘Сита думает, что та девушка, которая поет на телеэкране, красива. (Более точный перевод: ‘Какая девушка поет на телеэкране, Сита про ту думает, что она красивая’).

055-Mitrenina_OV:_Layout 1 18.05.2008 15:51 Page 359

(15b) *[jo vahã: rah-ta: hai]$_i$ mujhko [$_{NP}$ vo kaha:ni: [$_{RC}$jo Arundhati-ne us-
[Который там живет]$_i$ мне [$_{NP}$ та история [$_{RC}$ которую Арундхати
ke-baare-mẽ likh-ii]] pasand hai
того-про написал]] нравится
'Мне нравится история, которую Арундхати написал про того, кто там живет'.

Наличие барьера между главной и коррелятивной клаузой также говорит в пользу того, что коррелятивная клауза при переходе от D-структуры к Sструктуре перемещается влево.

Еще один формальный анализ коррелятива предполагает внешнее присоединение коррелятивной клаузы к главной клаузе.

(16) [$_{IP}$ [$_{CorCP}$ ... Rel-XP ...]$_i$ [$_{IP}$ ... Dem-XP$_i$ ...]]

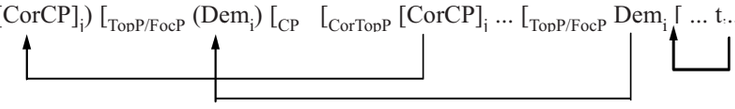
Р. Бхатт показал, что эта схема не годится для обычных коррелятивов в хинди, потому что в этом языке наблюдается барьер между коррелятивной и главной клаузой. Этот барьер свидетельствует о передвижении коррелятивной клаузы из состава главной клаузы в левую часть предложения. Однако схема (16), как отмечает в своей работе Р.Бхатт, подходит для анализа множественных коррелятивов в хинди.

Реконструкция не наблюдается в венгерских коррелятивных предложениях, как показала Анико Липтак [Lipták 2005: 13]:

(17) [$_{CorCP}$ Akit szeret Mari$_i$], azt meghívta pro$_i$ a buliba.
[$_{CorCP}$ Кого любит Мари$_i$], того приглашает pro$_i$ на вечеринку.
'Кого Мария любит, того она приглашает на вечеринку'.

А. Литпак предлагает свой вариант анализа подобных конструкций, в котором используется позиция коррелятивной темы Correlative Topic Phrase:

(18) [$_{CorTopP}$ ([CorCP]$_i$) [$_{TopP/FocP}$ (Dem$_i$) [$_{CP}$ [$_{CorTopP}$ [CorCP]$_i$ ... [$_{TopP/FocP}$ Dem$_i$ [... t$_i$....]]]]]



Но этот анализ подходит для венгерского, поскольку в этом языке фокус выделяется особо, а те группы, которые расположены перед ним, всегда являются топиком.

Учитывая эти схемы, мы рассмотрели, какой может быть структура коррелятивных предложений в русском языке.

Прежде всего, можно утверждать, что анализ (10a) для русских коррелятивов не подходит, поскольку в русском языке не возникает реконструкции в главном предложении:

(19) [Какая яхта$_k$ Мише$_i$ понравится]$_k$, ту$_k$ он$_i$ и покупает t$_k$.

След в конце этого предложения может соответствовать только определяемому слову в главной клаузе (*ту*), но без коррелятивной части, иначе местоимение *он* связывало бы кореферентное ему имя *Миша*, что сделало бы предложение таким же неприемлемым, как (11b) и (12b). Это значит, что в D-структуре коррелятивная клауза не может входить в состав главной клаузы, и анализ (10a) для русского языка не подходит.

Однако, как показал опрос информантов, носителей русского языка, барьеры наблюдаются и в русских коррелятивных конструкциях. В примерах (20a-c) барьеры отсутствуют, и эти предложения являются приемлемыми для большинства носителей русского языка, а в (21a-c) невозможно вынесение элемента из именной группы; большинство информантов признали предложения (21a-c) грамматически неправильными:

(20a) Кто честно работал, того ты пытаешься уволить, а [$_{CorCP}$**кто** воровал], [**ты хочешь, чтобы** [$_{IP}$я **того** опять взял к себе на работу]].

(20b) [$_{CorCP}$**Где**$_i$ его сон застанет]$_i$,[**я думаю, что** [$_{IP}$**там**$_i$ он и спит]].

(20c) [$_{CorCP}$**Какую** машину попросишь]$_i$, [я думаю, что [$_{IP}$он **такую** тебе и подарит]].

(21a) *[$_{CorCP}$**Кто** видит сквозь стены]$_i$, есть [$_{NP}$ больницы, где [$_{IP}$**таких** лечат бесплатно]]

(21b) *[$_{CorCP}$**Кто** плохо работал]$_i$, я выпустил [$_{NP}$указ, чтобы [$_{IP}$**того**$_i$ уволить]]].

(21c) *[$_{CorCP}$**Какой** город ему понравится]$_i$, бывали [$_{NP}$ случаи, когда [$_{IP}$он там останавливался на неделю или даже на две]]

Рассмотренные примеры показывают, что (a) в русском языке не происходит передвижения коррелятивной клаузы из главной части предложения; это значит, что в D-структуре коррелятивная клауза присоединяется к главной; (b) при этом коррелятивная клауза может отделяться от главной клаузы и перемещаться влево, как в предложениях (20a-c).

Вероятнее всего, расположение основных элементов коррелятивной конструкции связано с позициями топика и фокуса в русском языке. Наличие коррелятивной части всегда делает определяемое слово в главной клаузе старой информацией. Это подчеркивает и обязательное наличие указательного местоимения в определяемой группе в главной клаузе:

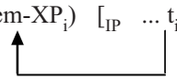
055-Mitrenina_OV:_Layout 1 18.05.2008 15:51 Page 360

(22a) Какую хотел, такую машину и купил.

(22b) *Какую хотел, машину и купил.

Предположительно структура коррелятива в русском языке может быть такой:

(23) $[_{TopP/FocP}\ [CorCP]_j\ [_{FocP}\ (Dem\text{-}XP_i)\ \ [_{IP}\ \ ...\ t_i...\ ]]]$



Интересно, что предположение о том, что коррелятивная часть в подобных конструкциях порождается как адъюнкт TopP, недавно было высказано также Е.Лютиковой [Лютикова 2008: 24]. Задача дальнейших исследований — более точно определить позицию коррелятивной клаузы и определяемого слова для коррелятивных предложений русского языка.

Список литературы

1. Зализняк А.А., Падучева Е.В. К типологии относительного предложения // Семиотика и информатика, М.: 1975 вып. 6.;

2. Лютикова Е. Загадки русских относительных предложений. Хэндаут доклада на конференции «Синтаксические структуры-2» 2-4 апреля 2008. М., РГГУ, 2008.

3. Тестелец Я.Г. Введение в общий синтаксис. М.: РГГУ, Наука, 2001.

4. Шведова Н.Ю. (ред.) Русская грамматика. Т. II. Синтаксис. М.: 1980.

5. Bhatt R. Locality in Correlatives // Natural Language and Linguistic Theory, Volume 21, Number 3, August 2003, pp. 485-541.

6. Dikken M. den. Comparative Correlatives Comparatively // Linguistic Inquiry, 2005. Volume 36:4, pp. 497 – 532.

7. Izvorski R. The syntax and semantics of correlative proforms // K. Kusumoto (ed.), Proceedings of NELS 26. Amherst, GLSA Publications, 1996.

8. Lipták A. Correlative Topicalization. uLCL, Leiden University ms. 2005.

9. Vries M de. The Syntax of Relativization. PhD thesis, University of Amsterdam, 2002.

10. Keenan E. Relative Clauses // T. Shopen, ed., Language typology and syntactic description, Cambridge, England: Cambridge University Press, 1985. Volume 2, pp. 141– 170.

056-Mitrofanova_O-Belik_V:_Layout 1 18.05.2008 15:51 Page 361

КОРПУСНОЕ ИССЛЕДОВАНИЕ СОЧЕТАЕМОСТНЫХ ПРЕДПОЧТЕНИЙ ЧАСТОТНЫХ ЛЕКСЕМ РУССКОГО ЯЗЫКА

CORPUS ANALYSIS OF SELECTIONAL PREFERENCES OF FREQUENT WORDS IN RUSSIAN

Митрофанова О.А. *(alkonost-om@yandex.ru),* ***Белик В.В.*** *(ogibbion14@pisem.net),*
Кадина В.В. *(veraiii@yandex.ru), Санкт-Петербургский государственный университет*

В докладе анализируются результаты исследования дистрибутивных свойств частотной лексики русского языка. Установлено решающее правило для выявления устойчивых сочетаний лексем с учётом коэффициента взаимной информации *MI*. Сочетаемостные предпочтения лексем определены в терминах морфологических классов и лексико-семантических признаков их синтагматических соседей.

1. Цели и задачи исследования

Информация о сочетаемостных предпочтениях слов, извлекаемая из корпусов текстов, играет важную роль при выполнении многих задач компьютерной лингвистики, среди которых автоматическая классификация лексики [Pekar, Staab 2003], разрешение неоднозначности [Resnik 1997], уточнение семантико-синтаксических моделей сочетаемости лексем в словарных базах данных [Азарова и др. 2005; Иорданская, Мельчук 2007; Ďurčo 2007], контрастивные исследования [Agirre et al. 2003] и пр.[1] Словари сочетаемости, построенные в результате компьютерной обработки больших корпусов текстов, представляют собой богатейший лингвистический ресурс [Гельбух и др. 2004]. В распоряжении лингвистов уже есть современные инструменты, предназначенные для исследования синтагматических свойств лексики и подключаемые непосредственно к корпусам. Например, существуют ресурсы (Sketch Engine[2]; Collocation Database, etc.[3]), которые позволяют не только ранжировать сочетания в соответствии с мерой их устойчивости, но также определять частеречную принадлежность, синтаксические и в некоторых случаях лексико-семантические признаки входящих в них слов [Lin 1999; Pala 2006]. Однако количественные критерии для выявления устойчивых сочетаний до сих пор недостаточно изучены.

В естественном языке существуют особые механизмы, которые регулируют комбинаторику лексических единиц текста на формальном и содержательном уровнях. Данные механизмы необходимо учитывать при моделировании понимания текста, в связи с этим функционирование сочетаний слов, тяготеющих к совместному употреблению, является объектом пристального внимания учёных в аспекте их статистической устойчивости, формальной и семантической связанности [Борисова 1995; Иорданская, Мельчук 2007; Ягунова 2006]. По всей видимости, анализ сочетаний слов с этих позиций даёт возможность детально исследовать их сочетаемостные предпочтения, модели взаимодействия их лексических значений, а также получить данные о контекстных маркерах значений.

Итак, цель обсуждаемого проекта – изучение дистрибутивных свойств частотной лексики в корпусах текстов русского языка, требующее решения ряда задач, среди которых:

• сбор и интерпретация данных о сочетаемости лексических единиц в корпусах текстов с учётом различных параметров (взаимное расположение элементов контекстов – правосторонние и левосторонние синтагматические соседи исследуемых лексем; веса элементов контекстов в зависимости от их позиции по отношению к исследуемым лексемам; ширина контекстного окна и пр.);

• получение количественных оценок силы связей лексических единиц в устойчивых сочетаниях; выявление и формулировка их сочетаемостных предпочтений с учётом морфологических классов и лексико-семантических признаков синтагматических соседей в контекстах.

[1] См. также материалы конференции CONTEXT: http://context-07.ruc.dk/CONTEXT07MainPage.html
[2] Sketch Engine: http://www.sketchengine.co.uk/; http://www.fi.muni.cz/~thomas/corpora/searches/index.htm
[3] Collocation Database, etc.: http://www.cs.ualberta.ca/~lindek/demos.htm

056-Mitrofanova_O-Belik_V:_Layout 1 18.05.2008 15:51 Page 362

2. Методика определения сочетаемостных предпочтений лексем

Сочетаемостные предпочтения лексемы *X* можно выявить, определив *{a, b, c…}* – множество её потенциальных синтагматических соседей в контекстах и упорядочив их с точки зрения различных признаков (например, принадлежность к ЛСГ – глаголы движения, интеллектуальной деятельности и пр., существительные – названия природных явлений, транспортных средств и пр., морфологический класс – глаголы, наречия, прилагательные, местоимения и пр., синтаксическая функция – актанты, сирконстанты, атрибуты, и пр.). Множество потенциальных синтагматических соседей лексемы *X* формируется в результате анализа выборочных совокупностей контекстов её употребления в корпусах текстов.

Количественный критерий предпочтительности синтагматических соседей *{a, b, c…}* для слова *X* может быть задан с учётом какого-либо коэффициента ассоциативной связи элементов в сочетаниях (например, в биграммах). В исследованиях применяются различные меры – *MI*, *T*, *Log-Likelihood*, *Z*, X^2, и пр. [Church, Hanks 1990; Evert, Krenn 2001]. В нашем случае был использован коэффициент взаимной информации *MI*, определяемый для биграмм типа *yX / Xy*, где $y \in \{a, b, c…\}$ – коллокат (левый или правый сосед) базовой лексемы *X*.

Коэффициент *MI* позволяет оценивать силу ассоциативной связи внутри сочетания слов (между лексемой *X* и её соседом *y*) на основе соотношения частоты встречаемости биграммы *f(X,y)* и независимых употреблений коллокатов *f(X)* и *f(y)*, с учётом объема корпуса *N*:

$$MI = \log_2 \left\{ \frac{N \bullet f(X,y)}{f(X) \bullet f(y)} \right\}$$

Чем выше значение коэффициента *MI*, тем более предпочтителен тот или иной синтагматический сосед *y* для лексемы *X* (и тем вероятнее, что *y* является маркером какого-либо из значений, закреплённых за *X*).

По сравнению с показателем частотности независимого употребления соседей лексемы *X*, коэффициент *MI* позволяет различать коллокаты с широкими сочетаемостными возможностями (которые могут оказаться высокочастотными и, вместе с тем, несущественными для лексемы *X*) и коллокаты, тяготеющие к употреблению в сочетаниях с лексемой *X* (и поэтому значимым образом характеризующие её сочетаемостные предпочтения).

Известно, что при извлечении биграмм из корпуса текстов с учётом значения *MI* удаётся выявить наибольшее число сочетаний, зарегистрированных в лексикографических источниках; доля биграмм со знаками пунктуации в экспериментах с *MI* оказывается существенно ниже, чем при использовании других мер, в частности, *T* и *Log-Likelihood* [Khokhlova 2007].

При формулировке сочетаемостных предпочтений слов предлагается использовать в качестве эвристики аппарат теории оптимальности, которая помогает смоделировать конкуренцию правил и ограничений, задействованных в построении языковых выражений на уровнях от фонологического до семантического [Blutner et al. 2006]. В зависимости от степени важности, эти правила и ограничения получают ранг. Чем важнее правило или ограничение, чем выше его ранг, тем серьёзнее его нарушение и тем менее «правильным» будет порождаемое языковое выражение. Правила и ограничения, имеющие низкий ранг, могут быть нарушены без ущерба для допустимости итогового языкового выражения. Иными словами, использование теории оптимальности в лингвистическом моделировании позволяет перейти от идеальных языковых структур к оптимальным (приемлемым в той или иной мере). С этих позиций можно установить иерархию приоритетов, существующих при выборе для лексемы *X* её синтагматических соседей с тем или иным лексическим значением, представляющих ту или иную часть речи, выполняющих ту или иную синтаксическую функцию.

3. Лингвистический материал, источники данных, исследовательские инструменты

Исследование проводится на материале наиболее частотных лексем русского языка, среди которых глаголы *идти, видеть, говорить, знать, сказать, есть, хотеть* и пр.; существительные *человек, год, рука, век, жизнь, друг, глаз* и пр.; прилагательные *близкий, далёкий, долгий, молодой, поздний, соседний, старший* и пр. Информация о сочетаемостных свойствах данных слов в дальнейшем использовалась при анализе контекстов их употребления в различных значениях. В ходе экспериментов осуществляется обработка лингвистических данных, содержащихся в ряде корпусных ресурсов: электронная библиотека М. Мошкова; корпус текстов русского языка Бокрёнок, применяемый на кафедре математической лингвистики СПбГУ; выборки типовых контекстов из Словаря русского языка С.И. Ожегова в формате базы данных Starling. Извлечение сочетаний слов производится с помощью сервиса поиска биграмм в лингвистическом ресурсе АОТ, где в качестве корпуса используется текстовая база электронной библиотеки М. Мошкова [Аверин 2006].[4] Данный сервис позволяет

[4] Сервис поиска биграмм в лингвистическом ресурсе АОТ: http://aot.ru/demo/bigrams.html

056-Mitrofanova_O-Belik_V:_Layout 1 18.05.2008 15:51 Page 363

получать списки биграмм с левосторонними / правосторонними коллокатами ключевого слова, упорядоченные по значению *MI*, по частоте биграммы или частотам коллокатов.

4. Формулировка решающего правила для выявления устойчивых сочетаний слов

Для содержательной обработки сочетаемостных данных необходимо сформулировать решающее правило, помогающее выявлять устойчивые сочетания, а главное, требуется определить соответствующее пороговое значение коэффициента взаимной информации *MI* в биграммах. Известно, что для сочетаний языковых единиц разных типов этот показатель должен подбираться индивидуально [Азарова и др. 2005]. Так, например, для английского языка с фиксированным порядком слов установлено пороговое значение *MI* = 3 [Church, Hanks 1990]. Можно допустить, что для русского языка эта величина будет несколько ниже, поскольку в русскоязычных текстах преобладает свободный порядок слов.

При определения порогового значения *MI* в качестве эвристики использовался метод минимального риска, или минимакса [Джонсон, Лион 1980: 433–435]. Суть метода заключается в том, что в процессе установления принадлежности какого-либо объекта к некоторому классу противопоставляются три типа решений: «попадание в цель» (правильное решение), «ложная тревога» (инородный объект ложно квалифицируется как входящий в класс) и «промах» (объект из класса не распознается как принадлежащий к классу). Правильные решения поощряются дополнительными очками или весами. Также производится взвешивание ошибок: менее серьёзные ошибки – «ложные тревоги» – получают меньший вес; более серьёзные ошибки – «промахи» – получают больший вес. В рассматриваемом случае трактовка устойчивого сочетания как неустойчивого следует считать «промахом», а обратную ситуацию – «ложной тревогой». Иллюстрацией «промаха» может служить игнорирование сочетаний со знаменательными словами, являющимися маркерами лексического значения базовой лексемы: например, *MI* (*говорить* + *язык*) = 0,777. «Ложные тревоги» чаще всего возникают в сочетаниях базовой лексемы и незнаменательных слов – местоимений, союзов, реже предлогов: например, *MI* (*говорить* + *я*) = 1,262.

При анализе биграмм было обнаружено, что оптимальное соотношение «попаданий в цель», «промахов» и «ложных тревог» достигается при *MI* = 1. В среднем, доля правильных решений составляет 87%, на десять «попаданий в цель» (вес «3») приходится один «промах» (вес «2») и две «ложных тревоги» (вес «1»). Изменение порогового значения приводит к снижению доли правильных решений и к увеличению доли ошибок. Таким образом, искомое решающее правило имеет следующий вид:

- если $MI \geq 1$, то сочетание слов считается устойчивым;
- если $MI < 1$, то сочетание слов оценивается как неустойчивое.

Расширенная версия данного решающего правила, учитывающая критерии для выявления связанных сочетаний различных типов (свободные / связанные, квазифраземы (коллокации) / фраземы, квазиидиомы / идиомы: согласно классификации, описанной в [Иорданская, Мельчук 2007]), приведена в работе [Митрофанова 2008].

Для верификации решающего правила было произведено сравнение результатов анализа биграмм, содержащих частотные лексемы русского языка, и информации об их синтагматических соседях, полученной в ходе ручной обработки представительных выборок из корпуса Бокрёнок [Митрофанова и др. 2006], а также типовых контекстов из Словаря русского языка С.И. Ожегова в формате базы данных Starling (СО) [Митрофанова, Крылов 2006]. Оказалось, что практически все синтагматические соседи, выявленные в контекстах из корпуса, встречаются в биграммах с $MI \geq 1$ (точнее, $MI \in [1, 3]$). Немногочисленным идиомам соответствуют биграммы с ещё более высоким значением *MI* (ср. *MI (речь* + *идти)* = 7,495; *MI (идти* + *вразрез)* = 9,466 и пр.)

Например, при интерпретации данных об употреблении существительного *человек* были обнаружены устойчивые сочетания, фигурирующие и в типовых контекстах из СО, и в контекстах из корпуса Бокрёнок, и в биграммах, при этом значение *MI* выше порогового:

MI (*молодой* + *человек*) = 6,339;
MI (*первобытный* + *человек*) = 5,645;
MI (*честный* +*человек*) = 5,212;
MI (*разумный* +*человек*) = 3,886;
MI (*хороший* + *человек*) = 2,536;
MI (*природа* + *человек*) = 2, 453;
MI (*честный* + *человек*) = 2,359;
MI (*жизнь* + *человек*) = 1,643;
MI (*отношение* + *человек*) = 1,112.

Также были рассмотрены другие сочетания существительного *человек* с левосторонними и правосторонними синтагматическими соседями, встретившиеся в контекстах из корпуса Бокрёнок и зарегистрированные в

056-Mitrofanova_O-Belik_V:_Layout 1 18.05.2008 15:51 Page 364

биграммах. Учитывалось положение соседей в сочетаниях, а также их тип с точки зрения решающего правила.

Левосторонние синтагматические соседи:

«попадания в цель»: прилагательные *здравомыслящий, порядочный, молодой, взрослый, умный, добрый, здоровый, простой, хороший, русский, живой, счастливый, близкий* и пр. количественные слова *миллиард, миллион, тысяча* и пр.; существительные *природа, сознание, судьба, жизнь, душа, сердце, мир* и пр.;

«ложные тревоги»: *этот, между, когда* и пр.; «промахи»: *любить, образ, имя* и пр.

Правосторонние синтагматические соседи:

«попадания в цель»: прилагательные *умный, добрый* и пр.; глаголы *обладать, иметь, жить, погибнуть, создать, начинать, заниматься, работать, уметь, пользоваться, сидеть, стоять, ходить, называть, считать, являться* и пр.;

«ложные тревоги»: *вообще, с, среди, ибо, то* и пр.; «промахи»: *нужно, хороший, молодой* и пр.

Тем самым, подтверждается предположение о том, что при выборе порогового значения *MI* = 1 удаётся учесть подавляющее большинство устойчивых сочетаний, при этом доля ошибок невелика.

5. Эксперименты по выявлению сочетаемостных предпочтений лексем в биграммах

Исследовательская процедура иллюстрируется на примере обработки биграмм с глаголом *идти* и прилагательным *далёкий*. Ниже приводятся фрагменты списков биграмм для данных слов, примеры их коллокатов в биграммах с $MI \geq 1$, сгруппированные на основе общности их морфологических и, где возможно, лексико-семантических признаков (таблицы 1–4). Данная информация была использована при формулировке и ранжировании сочетаемостных предпочтений изучаемых лексем.

5.1. Сочетаемость глагола идти

Биграммы, включающие глагол *идти* и левый контекст:

MI (*неторопливо* + *идти*) = 4,668;
MI (*смело* + *идти*) = 4,467;
MI (*поезд* + *идти*) = 4,278;
MI (*тропа* + *идти*) = 4,254;
MI (*надо* + *идти*) = 3,154;
MI (*решить* + *идти*) = 2,088;
MI (*мочь* + *идти*) = 1,770; и пр.

Морфологические классы	**Лексико-семантические признаки**	**Примеры**
наречия	скорость, время	*торопливо, неторопливо, медленно, быстро, долго* и пр.
	направление	*кругом, следом, впереди, навстречу, далее, далеко, куда, куда-то, некуда* и пр.
	эмоциональна оценка	*уверенно, смело, упорно* и пр.
существительные	средство передвижения	*караван, поезд, пароход* и пр.
	путь	*дорога, тропа* и пр.
	природное явление	*дождь, снег* и пр.
	сложное действие/процесс	*бой, разговор, торговля* и пр.
глаголы (в т.ч. предикативы)	∞	*мочь, продолжать, отказываться, молча, разрешить, решить, собираться, надо, пора* и пр.

Таблица 1. Группы левосторонних коллокатов глагола идти

Биграммы, включающие глагол *идти* и правый контекст:

MI (*идти* + *пешком*) = 7,240;
MI (*идти* + *ожесточённый*) = 5,475;
MI (*идти* + *далёкий*) = 4,642;
MI (*идти* + *спать*) = 3,860;
MI (*идти* + *отдыхать*) = 3,670;
MI (*идти* + *разговор*) = 2,175;
MI (*идти* + *волна*) = 1,218; и пр.

056-Mitrofanova_O-Belik_V:_Layout 1 18.05.2008 15:51 Page 365

Морфологические классы	**Лексико-семантические признаки**	**Примеры**
наречия	противопоставление	*напролом, наперекор* и пр.
	средство, способ	*пешком, босиком* и пр.
	направление	*рядом, впереди, следом, навстречу, домой, вдоль, параллельно, кругом, вперёд, прямо, напрямик, наверх, мимо, сюда, туда* и пр.
	положительная оценка	*нормально, гладко* и пр.
прилагательные (препозитивные определения в зависимых именных группах)	интенсивность	*ожесточённый, непрерывный* и пр.
	скорость	*медленный, быстрый* и пр.
	расстояние	*далёкий, близкий* и пр.
глаголы	∞	*завтракать, гулять, спать, отдыхать, идти* и пр.
существительные	природное явление	*дождь, пар, снег, волна* и пр.
	сложное действие/процесс	*подготовка, спор, бой, разговор* и пр.

Таблица 2. Группы правосторонних коллокатов глагола идти

5.2. Сочетаемость прилагательного далёкий

Биграммы, включающие прилагательное *далёкий* и левый контекст:
MI (*бесконечно* + *далёкий*) = 6,631;
MI (*донестись* + *далёкий*) = 3,823;
MI (*пробираться* + *далёкий*) = 3,804;
MI (*весьма* + *далёкий*) = 3,748;
MI (*немного* + *далёкий*) = 3,656;
MI (*вершина* + *далёкий*) = 2,279;
MI (*чужой* + *далёкий*) = 1,251; и пр.

Морфологические классы	**Лексико-семантические признаки**	**Примеры**
наречия	мера, степень	*бесконечно, страшно, весьма, немного, столь, настолько, очень, слишком, более, довольно* и пр.
прилагательные	∞	*невообразимый, далёкий, чужой, самый, такой, какой-нибудь* и пр.
глаголы	∞	*пробираться, донестись, послышаться, услышать* и пр.
существительные	путь	*путь*, *дорога* и пр.
	место	*страна, край, берег, вершина* и пр.

Таблица 3. Группы левосторонних коллокатов прилагательного далёкий

Биграммы, включающие прилагательное *далёкий* и правый контекст:
MI (*далёкий* + *прошлое*) = 6,592;
MI (*далёкий* + *предок*) = 6,181;
MI (*далёкий* + *звезда*) = 4,734;
MI (*далёкий* + *окраина*) = 4,223;
MI (*далёкий* + *галактика*) = 4,111;
MI (*далёкий* + *далёкий*) = 3,334;
MI (*далёкий* + *южный*) = 2,291; и пр.

Морфологические классы	**Лексико-семантические признаки**	**Примеры**
существительные	время	*прошлое, будущее, предок, потомок, детство, юность, древность*
	место	*родина, даль, край, окраина, страна, планета, звезда, галактика* и пр.
	сложное действие/процесс	*путешествие, плавание* и пр.
	природное явление, звук	*раскат, гром, эхо* и пр.
прилагательные	расстояние	*далёкий, прошлый* и пр.
	место, направление	*горный, северный, южный и пр.*

Таблица 4. Группы правосторонних коллокатов прилагательного далёкий

056-Mitrofanova_O-Belik_V:_Layout 1 18.05.2008 15:51 Page 366

5.3. Формулировка сочетаемостных предпочтений для лексем идти *и* далёкий

На основе информации о коллокатах лексем *идти* и *далёкий* удалось выявить морфологические модели типа *POS + X* / *X + POS* и ранжировать их в соответствии с наибольшими показателями *MI* в соответствующих группах биграмм. Рассматривались и другие способы ранжирования моделей (среднее геометрическое, мода), однако они не были достаточно эффективными.

Сочетаемостные предпочтения глагола *идти*:
ранг 1. *X + Adv*
ранг 2. *X + Adj*
ранг 3. *Adv + X*
ранг 4. *Noun + X*
ранг 5. *X + Verb*
ранг 6. *X + Noun*
ранг 7. *Verb + X*

Сочетаемостные предпочтения прилагательного *далёкий:*
ранг1. *X + Noun*
ранг 2. *Adv + X*
ранг 3. *X + Adj*
ранг 4. *Verb + X*
ранг 5. *Adj + X*
ранг 6. *Noun + X*

В ряде случаев оказалось возможным также сформулировать сочетаемостные предпочтения лексем в терминах лексико-семантических признаков их коллокатов, например, в сочетаниях типа *Adj + N*:

ранг 1. *Adj (далёкий) + N (время)*
ранг 2. *Adj (далёкий) + N (природное явление, звук)*
ранг 3. *Adj (далёкий) + N (место)*
ранг 4. *Adj (далёкий) + N (сложное действие/процесс)*
ранг 5. *N (место)+ Adj (далёкий)*
ранг 6. *N (путь) + Adj (далёкий)*

6. Итоги исследования и направления дальнейшей работы

В ходе исследования подтверждена возможность описания сочетаемостных предпочтений лексем на основе статистико-комбинаторных данных, извлекаемых из корпусов текстов, установлено решающее правило для выявления устойчивых сочетаний частотных лексем русского языка с учётом коэффициента взаимной информации MI. Сочетаемостные предпочтения рассматриваемых лексем определены в терминах морфологических классов и лексико-семантических признаков их коллокатов. Данная информация была использована в прикладных разработках: при анализе предикатно-аргументных структур в экспериментальном корпусе контекстов для частотных глаголов русского языка (кафедра математической лингвистики СПбГУ); в процессе создания словарных статей и подбора иллюстративных примеров употребления частотных прилагательных в проекте «Современный толковый словарь живого русского языка» (лаборатория компьютерной лексикографии СПбГУ).

В дальнейшем планируется:

- перевести процедуру выявления сочетаемостных предпочтений слов в полуавтоматический режим;
- произвести статистическую оценку эффективности метода выявления сочетаемостных предпочтений;
- дать лингвистическую и статистическую интерпретацию ошибочных решений;
- исследовать зависимость степени устойчивости сочетаний слов от синтаксической организации текста;
- применить разработанные описания сочетаемостных предпочтений лексем в процессе обучения компьютерного инструмента для разрешения лексико-семантической неоднозначности;
- осуществить эксперименты по разрешению лексико-семантической неоднозначности слов в русскоязычных текстах с учётом сочетаемостной информации.

Список литературы

1. Аверин А.Н. Разработка сервиса поиска биграмм // Труды Международной конференции «Корпусная лингвистика – 2006». СПб.: 2006. С. 5–15.

2. Азарова И.В., Синопальникова А.А., Смрж П. Представление устойчивых лексических сочетаний в компьютерном тезаурусе RussNet // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2005». М.: 2005. С. 11–17.

3. Борисова Е.Г. Коллокации. Что это такое и как их изучать. М., 1995.

4. Гельбух А.Ф., Сидоров Г.О., Эрнандес-Рубио Э., Чубукова М.В. Словари сочетаемости слов: какой метод составления лучше? // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2004». М.: 2004. URL: www.dialog-21.ru/Archive/2004/Gelbukh.pdf

056-Mitrofanova_O-Belik_V:_Layout 1 18.05.2008 15:51 Page 367

5. Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке. Т. 1. Методы обработки данных. М.: 1980.

6. Иорданская Л.Н., Мельчук И.А. Смысл и сочетаемость в словаре. М.: 2007.

7. Митрофанова О.А., Кадина В.В., Савицкий В.С. Словарь и корпус как источники данных о синтагматических связях лексических единиц // Труды Международной конференции «Корпусная лингвистика – 2006». СПб.: 2006. С. 271–281.

8. Митрофанова О.А., Крылов С.А. «Типовой» контекст: случайность или закономерность? // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2006». М.: 2006. С. 382–388.

9. Митрофанова О.А. О решающем правиле для определения устойчивости и связанности сочетаний слов // Четвёртая научно-практическая конференция «Прикладная лингвистика в науке и образовании». СПб.: 2008 [в печати].

10. Ягунова Е.В. Неоднословные целостности в словаре и корпусе // Труды Международной конференции «Корпусная лингвистика – 2006». СПб.: 2006. С. 395–412.

11. Agirre E., Aldezabal I., Pociello E. A Pilot Study of English Selectional Preferences and Their Cross-Lingual Compatibility with Basque // Text, Speech and Dialogue: 6th International Conference TSD–2003. Lecture Notes in Artificial Intelligence. Vol. 2807. Springer-Verlag: 2003. P. 12–19.

12. Blutner R., de Hoop H., Hendriks P. Optimal Communication. CSLI Lecture Notes. Vol. 177. Stanford: 2006.

13. Church K.W., Hanks P. Word Association Norms, Mutual Information, and Lexicography // Computational Linguistics. Vol. 16. 1990. P. 22–29.

14. Ďurčo P. Collocations in Slovak (Based on the Slovak National Corpus) // Computer Treatment of Slavic and East European Languages: 4th International Seminar. Bratislava: 2007. P. 43–50.

15. Evert S., Krenn B. Methods for the Qualitative Evaluation of Lexical Association Measures // Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. Toulouse: 2001. P. 188–195.

16. Khokhlova M. Collocations in Russian: Analysis of Association Measures // Computer Treatment of Slavic and East European Languages: 4th International Seminar. Bratislava: 2007. P. 96–103.

17. Lin D. Automatic Identification of Non-compositional Phrases // Proceedings of ACL–99. University of Maryland: 1999. P. 317–324.

18. Pala K. Word Sktеches and Semantic Roles // Труды Международной конференции «Корпусная лингвистика – 2006». СПб.: 2006. С. 307–317.

19. Pekar V., Staab S. Word Classification Based on Combined Measures of Distributional and Semantic Similarity // Proceedings of European Chapter of ACL–03, Research Notes Session. Budapest: 2003. P. 147–150.

20. Resnik P. Selectional Preference and Sense Disambiguation // Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? Washington: 1997. P. 52–57.

057-Mitrofanova_O-Panicheva_P:_Layout 1 18.05.2008 15:52 Page 368

СТАТИСТИЧЕСКОЕ РАЗРЕШЕНИЕ ЛЕКСИКО-СЕМАНТИЧЕСКОЙ НЕОДНОЗНАЧНОСТИ В КОНТЕКСТАХ ДЛЯ ПРЕДМЕТНЫХ ИМЁН СУЩЕСТВИТЕЛЬНЫХ

STATISTICAL WORD SENSE DISAMBIGUATION IN CONTEXTS FOR NAMES OF PHYSICAL OBJECTS

Митрофанова О.А. *(alkonost-om@yandex.ru),* ***Паничева П.В.*** *(ppolin@yandex.ru),*
Санкт-Петербургский государственный университет;
Ляшевская О.Н. *(olesar@mail.ru), Институт русского языка им. В.В. Виноградова РАН*

В докладе обсуждаются результаты экспериментов по автоматизации процесса разрешения лексико-семантической неоднозначности слов. Эмпирическим материалом исследования являются примеры употребления предметных имён, извлечённые из Национального корпуса русского языка. Оцениваются оптимальные условия разрешения неоднозначности с учётом двух факторов: лексического наполнения контекстов и лексико-семантической разметки контекстов.

1. Постановка проблемы, цели и задачи исследования

Неоднозначность, свойственная естественному языку и проявляющаяся на различных его уровнях, является серьёзным препятствием для компьютерного анализа текстов. Разрешение лексико-семантической неоднозначности (наряду с морфологической и синтаксической) имеет особую важность в подготовке корпусов текстов, используемых системами автоматического понимания естественного языка. Выполнение этой процедуры представляет наибольшую сложность и зачастую требует ручной обработки текстов лингвистами-экспертами, в распоряжении которых находятся обширные словарные картотеки. Качество ручного разрешения неоднозначности оценивается как высокое, вместе с тем, желательно снизить трудоёмкость данной задачи за счёт использования специализированных компьютерных инструментов.

Итак, целью настоящего исследования является автоматизация процесса разрешения лексико-семантической неоднозначности текстов, что требует выполнения ряда задач, среди которых:

• подготовка компьютерного инструмента автоматического разрешения лексико-семантической неоднозначности слов в контекстах;

• обработка экспериментальных выборок, содержащих неоднозначные контексты;

• определение оптимальных условий, при которых качество разрешения лексико-семантической неоднозначности слов в контексте было бы высоким.

2. Исследовательские методы

Известны достаточно эффективные методы разрешения лексико-семантической неоднозначности в полуавтоматическом или автоматическом режиме [WSD 2006].[1] Методы первого типа предполагают использование компьютерных тезаурусов (WordNet, FrameNet) и формальных онтологий в качестве источников информации о значениях слов. Методы второго типа основываются на статистических данных о контекстном окружении слов, позволяющем разграничивать их употребление в различных значениях.

Применительно к материалу русского языка опробованы оба типа методов. Использование мощного электронного лексикографического ресурса (РуТез, семантический словарь НКРЯ) обеспечивает высокий уровень разрешения лексико-семантической неоднозначности [Лукашевич, Чуйко 2007; Кустова и др. 2006; Шеманаева и др. 2007]. Если же есть необходимость обойтись без словарной поддержки (например, в том случае, если обрабатываются тексты больших объёмов, а их лексический состав не покрывается имеющимися в распоряжении исследователей словарями), то предпочтение следует отдать статистическим методам. Достаточно надёжным является разрешение лексико-семантической неоднозначности на основе сравнения дистрибуций частеречных тегов контекстного окружения слов [Азарова, Марина 2006] и на основе лексических маркеров контекстов [Кобрицов и др. 2005]. Допустимо совмещение тезаурусного и статистического подходов к

[1] См. также материалы конференции SENSEVAL (www.senseval.org) и библиографию работ по WSD в материалах Corpora List (http://listserv.linguistlist.org/cgi-bin/wa?A2=ind0512&L=corpora&D=1&F=&S=&P=2873).

057-Mitrofanova_O-Panicheva_P:_Layout 1 18.05.2008 15:52 Page 369

разрешению лексико-семантической неоднозначности с опорой на словарную информацию о моделях сочетаемости слов [Кобрицов и др. 2007]. Можно предположить, что не менее (а возможно даже более) эффективной окажется статистическое разрешение неоднозначности с учётом дистрибуций лексико-семантических тегов в контекстах. Таких исследований на материале корпусов русского языка до нынешнего времени не проводилось. Эксперименты подобного рода впервые осуществлены в рамках обсуждаемого проекта.

В целях изучения возможностей статистического разрешения лексико-семантической неоднозначности в русскоязычных текстах предлагается адаптировать компьютерный инструмент автоматической классификации лексики таким образом, чтобы производилось сравнение неоднозначных контекстов с эталонными контекстами, представляющими реализацию того или иного значения слова. Классификация контекстов может быть основана как на сходстве их лексического состава, так и на сходстве лексико-семантических тегов для контекстных элементов (при наличии соответствующей разметки корпуса текстов).

3. Экспериментальный материал

Эксперименты по разрешению лексико-семантической неоднозначности проводились на материале Национального корпуса русского языка (НКРЯ)[2]. Были запланированы эксперименты двух типов, предполагавшие снятие неоднозначности а) на основе лексических маркеров значений слов в контекстах (тег леммы) и б) на основе лексико-семантической разметки контекстов (теги первого значения слова).

В качестве тестовых лексем выбраны предметные имена существительные. Известна филиация значений данных слов, фиксируемая в лексико-семантической аннотации НКРЯ. При описании значений анализируемых лексем использовалась структура значений слов в [ТСРЯ 1992]. Каждому значению соответствует особая комбинация тегов, принятых в системе разметки НКРЯ[3] (см. таблицу 1). Для рассматриваемых слов были сформированы выборки контекстов, присутствующих в НКРЯ (объёмы выборок см. в таблице 1). Очевидно, что анализируемые лексемы отличаются количеством значений, характером развития полисемии/омонимии, сте-

Значения	**Лексико-семантическая аннотация**	**Примеры**	**Число контекстов из НКРЯ**
дом			**3000**, из них
m1a. Жилое (или для учреждения) здание	r:concr t:constr top:contain	*Дом – новостройка.*	1694
m1b. Свое жильё	r:concr t:space	*Брать работу на дом.*	95
m2. Семья, люди, живущие вместе, их хозяйство	r:concr t:group pt:set sc:hum	*Мы знакомы домами.*	72
m3. Место, где живут люди, объединенные общими интересами, условиями существования	r:concr t:space der:shift der:metaph	*Общеевропейский дом.*	4
m4. Учреждение, заведение, обслуживающее какие-нибудь общественные нужды	r:concr t:org	*Дом культуры.*	292
m5. Династия, род	r:concr pt:set sc:hum	*Дом Романовых.*	1
диффузные значения *m1a/m1b, m1a/m2, m1b/m2* и пр.			842
орган			**834**, из них
m1. Клавишный духовой музыкальный инструмент, состоящий из труб, в к-рые нагнетается воздух	r:concr t:tool:mus	*Играть на органе.*	27
m2. Часть организма, имеющая определённое строение и специальное назначение	r:concr pt:partb pc:hum pc:animal hi:class	*Орган слуха.*	130
m2a. Орудие, средство	r:concr der:shift dt:partb	*Печать – активный орган пропаганды.*	9
m3. Государственное или общественное учреждение, организация	r:concr t:org hi:class	*Органы здравоохранения.*	660
m4. Печатное издание, принадлежащее какой-н. партии, организации, учреждению	r:concr t:media hi:class	*Академический орган.*	8
лук			**2200**, из них
m1. Огородное или дикорастущее растение сем. лилейных с острым вкусом луковицы и съедобными трубчатыми листьями	r:concr t:plant t:fruit t:food pt:aggr	*Репчатый лук.*	1600
m2. Ручное оружие для метания стрел в виде пружинящей дуги, стянутой тетивой	r:concr t:tool:weapon top:arc	*Стрельба из лука.*	600

Таблица 1. Филиация значений слов дом, орган, лук

[2] Публикации по НКРЯ: http://www.ruscorpora.ru/corpora-biblio.html

[3] Подробное описание системы тегов: http://www.ruscorpora.ru/corpora-sem.html

057-Mitrofanova_O-Panicheva_P:_Layout 1 18.05.2008 15:52 Page 370

пенью связанности значений между собой. Необходимо отметить, что в рамках данного исследования используется трактовка неоднозначности, принятая в компьютерной лингвистике и допускающая условное приравнивание омонимичных коррелятов к многозначным словам [Рахилина и др. 2006]. Поэтому указанный материал для экспериментов по автоматическому разрешению неоднозначности является репрезентативным и позволяет получить результаты, соотносимые с разными условиями разрешения лексико-семантической неоднозначности.

Эксперименты по разрешению лексико-семантической неоднозначности проводились только для значений, представленных в НКРЯ достаточным количеством контекстов (например, из рассмотрения были исключены значения *m3* и *m5* для слова *дом*, значения *m2a* и *m4* для слова *орган*).

Было учтено, что в ряде контекстов регистрируется диффузность значений исследуемых лексем: например, *дом – m1a* (строение) vs. *дом – m1b* (личное пространство, которое часто физически оказывается вовсе не домом, а комнатой или квартирой, ср. отыменные наречия *дома*, *домой*). Подобные контексты были проанализированы отдельно.

4. Постановка экспериментов

Разрешение лексико-семантической неоднозначности слов в корпусе рассматривается как задача распознавания образов. В качестве экспериментальной выборки используется набор контекстов, в которые вручную введены лексико-семантические теги, соответствующие значениям исследуемых лексем. Из экспериментальной выборки контекстов для той или иной лексемы автоматически формируются образы – эталонные классы контекстов, иллюстрирующие употребление слова в каком-либо одном значении. В образ попадают контексты, отобранные случайно. Оставшиеся тестовые контексты (все или часть из них) автоматически сравниваются с образами и распределяются по группам в соответствии со значениями, в этом случае априорная лексико-семантическая информация об исследуемых лексемах не используется: значение лексемы определяется автоматически. Тем самым, разрешение неоднозначности предполагает автоматическую классификацию контекстов употребления лексемы в разных значениях. Данная процедура требует представления экспериментальной выборки как векторного пространства, где каждый контекст преобразуется в вектор. Близость контекста употребления слова в каком-либо значении к тому или иному образу оценивается с помощью трёх мер расстояния: меры Евклида (*Eucl*), меры Хемминга (*Hm*) и значения косинуса угла между контекстными векторами (*Cos*). Данные меры имеют некоторые особенности. Если мера Хемминга линейна (и она аккумулирует разницы по координатам для двух точек), то мера Евклида отражает квадратичную зависимость расстояния между точками от разниц по их координатам (она аккумулирует квадраты разниц по координатам). В обоих случаях на результат влияют как раз большие разницы, это влияние слабее для меры Хемминга и сильнее для меры Евклида. В отличие от меры Евклида и меры Хемминга, мера косинуса менее чувствительна к большим разницам по отдельным координатам и не зависит от длин векторов.

Для исследуемых слов была проведена серия экспериментов с различными по объёму эталонными классами и тестовыми выборками контекстов, с изменением меры близости, с опорой на лексические маркеры значения в контексте либо на лексико-семантические теги. Во всех экспериментах объём контекста не ограничивался каким-либо окном. Результаты автоматической классификации контекстов сравнивались с результатами ручной разметки значений слов в контекстах.

5. Компьютерное обеспечение экспериментов

В экспериментах использовался компьютерный инструмент автоматической классификации лексики [Митрофанова и др. 2007], адаптированный для разрешения неоднозначности слов в контексте. Реализован алгоритм классификации с учителем. Программное обеспечение разработано П.В. Паничевой на языке Python. В ходе работы программы производятся следующие процедуры.

Во-первых, производится подготовительная обработка экспериментальных выборок контекстов. В выборке определяются те контексты, в которых значение лексемы может быть идентифицировано однозначно. Вычисляется количество имеющихся контекстов для каждого из значений исследуемой лексемы. Для значений с достаточным числом контекстов случайным образом формируется тестовая выборка и не пересекающаяся с ней обучающая выборка (эталонный класс). Для дальнейшей работы программы необходимо, чтобы для каждого значения были сформированы два файла, в которых приведены тестовая выборка и эталонный класс.

Во-вторых, осуществляется процесс машинного обучения. Для исследуемых значений программа производит обработку файла с эталонными классами контекстов, в ходе которой формируется образ значения. Из эталонных контекстов извлекается лексическая информация, тем самым, в образ значения включаются все лексемы, встретившиеся в эталонных контекстах, с учётом частоты их встречаемости. На выходе процедуры

057-Mitrofanova_O-Panicheva_P:_Layout 1 18.05.2008 15:52 Page 371

формируется статистические образы значений анализируемого слова, представленные словарём, в котором указаны лексемы и их относительная частота. Таким образом, если обучающая выборка для одного из значений слова *лук* составляла бы 100 контекстов, и в них 50 раз встретилась лексема *резать* и 30 раз встретилась лексема *морковь*, то в статистическом образе этого значения глагол *резать* имел бы показатель частотности 0,5, а существительное *морковь* – 0,3. Итак, образ значения можно рассматривать как вектор в векторном пространстве, координаты которого определяются частотными показателями соответствующих лексем, встретившихся в обучающей выборке контекстов для этого значения. В экспериментах с учётом лексико-семантической информации статистический образ формируется аналогичным путём, однако координатами в векторном пространстве служат не слова, а лексико-семантические теги слов, выступающих в качестве контекстного окружения исследуемых лексем.

Далее программа, прошедшая обучение, обрабатывает тестовые выборки контекстов. Для этого каждый контекст также рассматривается как вектор в векторном пространстве, и вычисляется мера расстояния данного контекста по отношению к векторам, представляющим образы значений. Выбирается образ значения, который оказывается наиболее близким к образу анализируемого контекста, в итоге, этому контексту присваивается соответствующее значение. При проверке результатов классификации для каждого из значений вычисляется количество правильных решений – тех случаев, когда автоматическая оценка значения, реализованного в контексте, совпадает со значением, назначенным вручную и отражённым в лексико-семантических тегах исследуемой лексемы.

6. Результаты экспериментов по автоматическому разрешению лексико-семантической неоднозначности слов в контекстах

6.1. Иллюстрация результатов компьютерной обработки контекстов

В ходе экспериментов обрабатываемым неоднозначным контекстам для предметных имён существительных автоматически приписывалось то или иное значение. Так, в таблице 2 приведены некоторые примеры анализа контекстов слова *дом*.

Контексты (в квадратных скобках указан номер контекста в корпусе)	**Исходное значение**	**Распознанное значение**	***Cos***
[649] Я помню всю эту чепуху детства, потери, находки, то, как я страдал из-за него, когда он не хотел меня ждать и шёл в школу с другим, и то, как передвигали *дом* с аптекой, и ещё то, что во дворах всегда был сырой воздух, пахло рекой, и запах реки был в комнатах, особенно в большой отцовской, и, когда шёл трамвай по мосту, металлическое бренчание и лязг колёс были слышны далеко.	*m1a*	*m1a*	0,650
[3004] Уже два года, как Таня ушла *из дому* и жила по разным местам, у новых приятелей, – то в мастерской знакомого художника на Шаболовке, то на пустующей зимней даче чьих-то родственников под Звенигородом, то в служебной квартире подружки, работавшей техником-смотрителем на Молчановке...	*m1b*	*m1b*	0,438
[957] Все подъезды в этом *доме* – со двора.	*m1a*	*m4*	0,288
[2130] Домишко рядом с *домом* подполковника.	*m1a*	*m2*	0,099
[3042] Пришлось Анну вернуть в *дом*, вскоре и Катю поселили.	*m1b*	*m4*	0,410

Таблица 2. Примеры компьютерной обработки контекстов употребления слова дом

Примеры [649] и [3004] проанализированы верно, тогда как примеры [957], [2130] и [3042] интерпретируются неточно. Вероятно, ошибочные решения связаны с недостаточностью контекстного окружения для идентификации значений.

Результаты автоматического разрешения неоднозначности дополняются информацией о контекстных маркерах лексических значений исследуемых слов в контекстах (см., например, таблицу 3).

Значения	**Лексические маркеры**
m2. Часть организма…	*порок, врождённый…*
m3. …Учреждение, организация…	*учреждение, самоуправление, начальник, местный, правоохранительный…*

Таблица 3. Примеры лексических маркеров значений слова орган *в контекстах*

057-Mitrofanova_O-Panicheva_P:_Layout 1 18.05.2008 15:52 Page 372

6.2. Оптимальные условия автоматического разрешения лексико-семантической неоднозначности слов в контекстах

Данные, полученные в процессе исследования, свидетельствуют о следующих фактах.

Во-первых, наилучшие результаты разрешения лексико-семантической неоднозначности на основе лексических маркеров (в среднем 85% правильных решений, в отдельных случаях до 95% правильных решений) могут быть получены при использовании в качестве меры расстояния значения косинуса угла между контекстными векторами (см. таблицу 4) .

Мера	***Eucl***	***Hm***	***Cos***
Точность *(p)*	0,45	0,65	0,85

Таблица 4. Точность результатов автоматического разрешения лексико-семантической неоднозначности слов в контекстах с использованием различных мер

Во-вторых, успешность разрешения лексико-семантической неоднозначности находится в прямой зависимости от частотности контекстов с тем или иным значением слова в экспериментальной выборке. Частотность значения сказывается на чёткости формируемого эталонного класса. Эталонные классы для частотных значений являются более чёткими, чем классы для значений с умеренной частотой. Так, для слова *орган* высокочастотное значение *m3* распознаётся лучше, чем низкочастотное значение *m1* и значение *m2* с умеренной частотой. По всей видимости, хороших результатов распознавания можно достигнуть при наличии не менее 100 контекстов употребления слова в экспериментальной выборке.

В-третьих, изменение объёма эталонного класса (S = 15, 55, 75, 100, 200, 500, … полная выборка за исключением тестовых контекстов) также оказывает существенное влияние на качество разрешения лексико-семантической неоднозначности. При предельных объёмах эталонных классов качество распознавания оказывается низким, поскольку в эталонном классе малого объёма недостаточно контекстов для фиксации признаков употребления слова в том или ином значении, а в максимально широком эталонном классе велика доля случайных признаков, не сопряжённых с конкретным значением (см., например, таблицы 5 и 6).

Объём эталонного класса *(S)*	**Точность *(p)***	**Объём эталонного класса *(S)***	**Точность *(p)***	**Объём эталонного класса *(S)***	**Точность *(p)***
15	0,63	75	0,77	200	0,56
55	0,80	100	0,8	полная выборка	0,77

Таблица 5. Точность результатов автоматического разрешения лексико-семантической неоднозначности слова орган *в контекстах с использованием меры* Cos *и с учётом объёма эталонного класса*

Объём эталонного класса *(S)*	**Точность *(p)***	**Объём эталонного класса *(S)***	**Точность *(p)***	**Объём эталонного класса *(S)***	**Точность *(p)***
100	0,78	500	0,83	полная выборка	0,73

Таблица 6. Точность результатов автоматического разрешения лексико-семантической неоднозначности слова лук *в контекстах с использованием меры* Cos *и с учётом объёма эталонного класса*

6.3. Сравнение результатов автоматического разрешения лексико-семантической неоднозначности на основе лексических маркеров и лексико-семантических тегов

Была проведена серия экспериментов для сравнения эффективности автоматического разрешения лексико-семантической неоднозначности слов на основе лексических маркеров, выявляемых в их контекстах, и лексико-семантических тегов их контекстного окружения. Например, в таблице 7 приведены некоторые контексты, иллюстрирующие употребление слова *лук* в значениях *m1* и *m2*, а также результаты их компьютерной обработки в двух режимах (объём тестовых выборок – 20 контекстов, объём эталонных классов – 500 контекстов, мера *Cos*).

057-Mitrofanova_O-Panicheva_P:_Layout 1 18.05.2008 15:52 Page 373

Контексты (в квадратных скобках указан номер контекста в корпусе)	**Распознавание на основе лексических маркеров**		**Распознавание на основе лексико-семантических тегов**	
	Распознанное значение	***Cos***	**Распознанное значение**	***Cos***
исходное значение m1				
[2379] Помню хлеб с изюмом, с *луком*, с какими-то кореньями.	*m1*	0,572	*m1*	0,786
[1578] Щавель –300 г, огурцы – 50 г, *лук* зелёный – 30 г, яйца – 1 шт., сметана – 30 г, сахар – 10 г, укроп.	*m1*	0,653	*m1*	0,569
[193] Начинают принимать *лук*, капусту – гляди в оба глаза.	*m2*	0,502	*m1*	0,514
исходное значение m2				
[235] Одни тугие *луки*, над которыми несколько человек справиться не могли, «играючи» натягивали, другие толстенные железные полосы вокруг шеи врага скручивали, третьи возы через броды на себе перетаскивали, ядра через самые широкие реки запросто перебрасывали.	*m2*	0,533	*m2*	0,550
[1120] Знаешь, есть восточное присловье, что, если человек стреляет из *лука*, он никогда не попадет в мишень, если стрела не пробьет одновременно его сердце.	*m2*	0,543	*m2*	0,538
[1863] Не имев совершенного успеха в намерении взбунтовать тушинский стан и боясь мести гетмана, Марина, в одежде воина, с *луком* и тулом за плечами, [11 февраля] ночью, в трескучий мороз ускакала верхом к мужу, провождаемая только слугою и служанкою.	*m1*	0,507	*m2*	0,609

Таблица 7. Примеры компьютерной обработки контекстов употребления слова лук

Оценки точности автоматического разрешения лексико-семантической неоднозначности при заданных условиях приведены в таблице 8.

	Точность *(p)*		**Среднее *($p_{ср}$)***
	лук *(m1)*	**лук *(m2)***	
Распознавание на основе лексических маркеров	0,75	0,9	0,83
Распознавание на основе лексико-семантических тегов	0,75	0,95	0,85

Таблица 8. Точность результатов автоматического разрешения лексико-семантической неоднозначности слова лук *в контекстах на основе лексических маркеров и лексико-семантических тегов*

В подавляющем большинстве случаев распознавание на основе лексических маркеров и на основе лексико-семантических тегов приводит к одинаково правильным решениям (см. примеры [2379], [1578], [235], [1120] в таблице 7). Вместе с тем, результаты разрешения лексико-семантической неоднозначности по тегам часто оказываются лучше, чем результаты, полученные при использовании лексических маркеров (ср. значения меры косинуса для примеров [2379] и [235]). Были зарегистрированы контексты, показывающие незначительное снижение значения меры косинуса (ср. примеры [1578] и [1120]), однако это не влияет на качество распознавания при переходе от лексических маркеров к тегам. Важно, что в ходе анализа экспериментальных данных удалось получить подтверждение гипотезы о том, что при разрешении неоднозначности на основе лексико-семантических тегов удаётся улучшить результаты идентификации значений слов в контексте и избежать ошибочных решений (см. примеры [193] и [1863]). Среди причин, вызывающих неудачи при разрешении лексико-семантической неоднозначности, можно указать недостаточность (вплоть до полного отсутствия) диагностических маркеров значения в чрезмерно коротких контекстах (см. пример [193]) или, наоборот, в слишком широких контекстах (см. контекст [1863]). Как правило, значение меры косинуса в этих случаях удерживается около показателя 0,5. Возможный путь корректировки результатов автоматического анализа связан с дополнительным использованием других мер расстояния.

057-Mitrofanova_O-Panicheva_P:_Layout 1 18.05.2008 15:52 Page 374

6.4. Анализ контекстов с диффузными значениями

Наряду с экспериментами по автоматической обработке потенциально однозначных контекстов употребления слов было произведено разрешение лексико-семантической неоднозначности в контекстах с диффузными значениями, а также сравнение результатов ручного и компьютерного анализа. В таблице 9 приведены примеры некоторых диффузных контекстов слова *дом*, указывающие на возможность выбора доминирующего значения в паре по итогам компьютерного анализа.

Контексты (в квадратных скобках указан номер контекста в корпусе)	Диффузные значения	Распознанное значение	**Cos**
[337] А в *доме* у Ёжика топилась печь, потрескивал в печи огонь, а сам Ёжик сидел на полу у печки, помаргивая, глядел на пламя и радовался.	*m1a*/*m1b*	*m1a*	0,429
[2983] Семён на портфель и не взглянул, а заточку аккуратно обтёр кухонной тряпкой, предусмотрительно им захваченной *из дому*, засунул инструмент в рукав, под часовой ремень, и вышел из двора той новой походкой, негнущейся и манекенной, которая образовалась у него после больничного излечения...	*m1a*/*m1b*	*m1b*	0,541
[3214] Родственники у Ливии все как один люди практичные, богатые и важные, хоть и не без вывертов; кажется, единственный человек, который уважает её в этом *доме*, – это ее дворецкий, Трефль.	*m1b*/*m2*	*m2*	0,452

Таблица 9. Примеры компьютерной обработки сложных случаев употребления слова дом *в контекстах*

В дальнейшем условия эксперимента были изменены, дополнительно сформированы эталонные классы для диффузных значений типа *m1a*/*m1b*, *m1a*/*m2*, *m1b*/*m2* и пр.

7. Выводы и перспективы развития исследования

В результате исследования была проведена модернизация компьютерного инструмента автоматической классификации лексики и введение специализированного режима его работы, позволяющего автоматически классифицировать неоднозначные контексты употребления слов в соответствии с присущими им значениями. Был реализован алгоритм классификации объектов с учителем и процедуры автоматической обработки контекстов с опорой на лексическое наполнение контекстов, а также с учётом лексико-семантических тегов, приписываемых контекстному окружению слов.

Были проведены серии экспериментов по автоматическому разрешению неоднозначности контекстов употребления предметных имён существительных с различной семантической структурой. Данные слова характеризуются разным числом значений, отличающихся по частотности и по степени самостоятельности. Это позволило получить обширные экспериментальные данные на русскоязычном материале и оценить оптимальные условия, обеспечивающие достаточно высокое качество разрешения семантической неоднозначности слов в контекстах (от 85% и выше).

Оптимальными можно признать следующие условия разрешения лексико-семантической неоднозначности слов в контекстах:

• высокий объём экспериментальной выборки;

• наличие в выборке не менее 100 контекстов употребления слова в исследуемом значении;

• объём эталонного класса около 500 контекстов;

• оценка близости контекстов к эталонному классу с использованием значения косинуса угла между контекстными векторами;

• возможность снятия неоднозначности на основе лексических маркеров значения слова в контексте либо на основе лексико-семантических тегов его контекстного окружения.

В ходе экспериментов нашла подтверждение гипотеза о большей эффективности разрешения лексико-семантической неоднозначности с опорой на лексико-семантическую разметку корпуса текстов.

Продолжение исследования предполагает проведение экспериментов по разрешению семантической неоднозначности:

• на обширном корпусном материале (увеличение экспериментальной группы лексем, использование большеобъёмных экспериментальных выборок контекстов из корпуса);

• с оценкой контекста на основе комбинированных признаков (например, с учётом как лексических, так и лексико-семантических данных, с вычислением оптимальных весовых коэффициентов в контекстах и пр.);

057-Mitrofanova_O-Panicheva_P:_Layout 1 18.05.2008 15:52 Page 375

• с изменением ширины контекстного окна (в предыдущих экспериментах рассматривались контексты в полном объёме, предлагается сужать границы контекстов и варьировать протяжённость обрабатываемых фрагментов контекстов);

• с детальным анализом диффузных контекстов употребления лексем в сопряжённых значениях (определение доминирующего значения: например, *стакан с водой* (*стакан* – «вместилище») vs. *стакан воды* (*стакан* – «мера+вместилище»);

• с проверкой ряда статистических гипотез об условиях разрешения лексико-семантической неоднозначности лексем в корпусах текстов.

Список литературы

1. Азарова И.В., Марина А.С. Автоматизированная классификация контекстов при подготовке данных для компьютерного тезауруса RussNet // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2006». М.: 2006. С. 13–17.

2. Кобрицов Б.П., Ляшевская О.Н., Шеманаева О.Ю. Снятие лексико-семантической омонимии в новостных и газетно-журнальных текстах: поверхностные фильтры и статистическая оценка // Интернет–математика 2005: Автоматическая обработка веб-данных. М.: 2005. С. 38–57.

3. Кобрицов Б.П., Ляшевская О.Н., Толдова С.Ю. Снятие семантической многозначности глаголов с использованием моделей управления, извлечённых из электронных толковых словарей // URL: http://download.yandex.ru/IMAT2007/kobricov.pdf

4. Кустова Г.И., Рахилина Е.В., Ляшевская О.Н., Шеманаева О.Ю. Семантическая разметка и семантические фильтры для Национального корпуса русского языка // Труды международной конференции «Корпусная лингвистика–2006». СПб.: 2006. С. 209–218.

5. Лукашевич Н.В., Чуйко Д.С. Автоматическое разрешение лексической многозначности на базе тезаурусных знаний // Интернет–математика 2007: Сборник работ участников конкурса. Екатеринбург: 2007. С. 108–117.

6. Митрофанова О.А., Мухин А.С., Паничева П.В. Автоматическая классификация лексики в русскоязычных текстах на основе латентного семантического анализа // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2007». М.: 2007. С. 413–421.

7. Рахилина Е.В., Кобрицов Б.П., Кустова Г.И., Ляшевская О.Н., Шеманаева О.Ю. Многозначность как прикладная проблема: лексико-семантическая разметка в Национальном корпусе русского языка // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2006». М.: 2006. С. 445–450.

8. ТСРЯ – Ожегов С.И., Шведова Н.Ю. Толковый словарь русского языка. М., 1992.

9. Шеманаева О.Ю., Кустова Г.И., Ляшевская О.Н., Рахилина Е.В. Семантические фильтры для разрешения многозначности в национальном корпусе // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2007». М.: 2007. С. 582–587.

10. WSD – Word Sense Disambiguation: Algorithms and Applications / Eds. E. Agirre, Ph. Edmonds. Springer: 2006.

058-Mihailov_M:_Layout 1 18.05.2008 15:53 Page 376

КОРПУС УСТНЫХ ПЕРЕВОДОВ КАК НОВЫЙ ТИП КОРПУСА ТЕКСТОВ

THE INTERPRETING CORPUS AS A NEW TYPE OF TEXT CORPUS

Михайлов М.Н. *(mihail.mihailov@uta.fi),* ***Исолахти Н.Б.*** *(nina.isolahti@uta.fi)*
Тамперский университет, Тампере, Финляндия

Я теперь пишу не роман, а роман в стихах – дьявольская разница.

(из письма А.С. Пушкина П.А. Вяземскому)

Обсуждается разработка корпуса устных переводов на примере судебного дискурса. Корпус сочетает в себе корпус устной речи и параллельный корпус. Разметка должна содержать коммуникативную, просодическую и экстралингвистическую информацию. Данные корпуса интересны для междисциплинарных исследований.

1. На пути к корпусу дискурса

За последние десятилетия корпусная лингвистика сделала гигантский скачок. Корпуса текстов из «роскоши« стали «средством передвижения»: на сегодняшний день редкое исследование по лингвистике обходится без привлечения корпуса текстов. Тем не менее нельзя не отметить, что развитие это идет неравномерно: в первую очередь растут одноязычные общеязыковые корпуса текстов (такие, как British National Corpus, Национальный корпус русского языка и т.п.) при явном отставании параллельных, диахронических корпусов и т.п.

Развитие корпусов устной речи также, по понятным причинам, идет более медленными темпами. Устная составляющая больших корпусов текстов несопоставимо мала по сравнению с письменной, например, устная речь составляет лишь 3,9% от общего объема Русского национального корпуса (http://www.ruscorpora.ru/corpora-stat.html, 08.04.2008). Главной проблемой при создании корпусов устной речи являются крайне медленные темпы ввода данных, поскольку какая-либо автоматизация на данный момент отсутствует. Не менее сложной оказывается и разметка корпуса. Из коллекций транскрибированной[1] устной речи лишь монологи теоретически можно организовать по тем же принципам, что и письменные тексты. Диалоги нельзя хранить в формате «текстовых» корпусов текстов. В противном случае неизбежны очень существенные информационные потери и заметные неудобства.

В корпусе текстов, о котором пойдет речь в нашей статье, представлен еще более неудобный для хранения и аннотирования материал: дискурс с устным переводом. В ходе такого дискурса имеет место обмен репликами на двух языках, часть которых (но не обязательно все!) переводится на другой язык, причем перевод может происходить в обе стороны и с наложением во времени (то есть в некоторых случаях последовательный перевод переходит в синхронный).

Корпус текстов судебных переводов становится в некоторой степени междисциплинарным продуктом. Судебный дискурс интересен для исследователей как особый тип речевого общения. Исследования проводятся в разных направлениях: изучается, например, влияние институционального контекста на деятельность участников дискурса, роли участников дискурса, их интерактивная деятельность, взаимодействие и взаимовлияние, речевое выражение иерархии и конфликта взаимоотношений (напр., Красовская 2006; Välikoski 2004). Устный перевод вносит в коммуникацию дополнительный компонент: роли участников коммуникативной ситуации осложняются межкультурными и межъязыковыми различиями.

Судебный перевод остается малоизученным элементом институционального дискурса. Работы последних лет посвящены изучению роли переводчика в судебном процессе, профессионализма, техники судебного перевода (González, Vásquez & Mikkelson 1991; Mikkelson 2000), однако появились и работы, рассматривающие судебный перевод как часть судебного дискурса (Hale 2004; Moeketsi 1999).

Аутентичный материал, показывающий все так, «как это было», с паузами, исправлениями, повторами и

[1] В данной статье транскрипция понимается в широком смысле как способ передачи устной речи с помощью письменных знаков.

058-Mihailov_M:_Layout 1 18.05.2008 15:53 Page 377

оговорками, представляет для таких исследований огромную ценность. На кафедре перевода русского языка Института современных языков и переводоведения Тамперского университета создается корпус **CoInCoUT** (Court Interpreting Corpus at the University of Tampere). Материал представляет собой собрание аудиозаписей допросов подсудимых и свидетелей с участием переводчиков языковой пары финский-русский. Записи транскрибируются и аннотируются. На настоящий момент в корпусе имеются аудиозаписи 8 процессов, в которых работало 3 разных переводчика, общая продолжительность записей около 14 часов. Корпус постоянно пополняется.[2]

2. Архитектура корпуса устного перевода

Корпус текстов организован в формате базы данных Postgresql с доступом к ней через веб-интерфейс, написанный на языке PHP 5, и размещен на Linux-сервере Института современных языков и переводоведения mustikka.uta.fi (доступ на сервер ограничен). Тексты аудиозаписей хранятся в виде таблицы, каждая запись которой представляет собой текстоформу, просодический элемент или элемент разметки («тэг»). В связанной с ней таблице хранится информация о каждой реплике: говорящий, момент начала и конца говорения, язык, реплика-«хозяин» (например, если реплика является переводом другой реплики). Леммы текстоформ помещены в отдельные две таблицы, связанные с таблицей текстов — таблицу русских лемм и таблицу финских лемм.[3] Еще одна таблица служит для хранения информации о судебных заседаниях, с которых были получены записи[4].

Таким образом, путем организации корпуса в виде набора связанных таблиц, удается придать некую форму этому нагромождению реплик на разных языках, произносимых разными людьми. Вообще говоря, представляется, что данные такого типа можно на сегодняшний день хранить без существенных потерь только в виде SQL базы данных либо в формате XML.

Для выполнения различных запросов к корпусу текстов написано большое количество различных утилит, объединенных в веб-интерфейс. С помощью него пользователь может получать конкордансы, словники, наборы коллокаций и т.п. Поскольку корпус создан в первую очередь для исследования дискурса и процесса перевода, то и конкорданс в таком корпусе будет несколько отличаться от конкорданса в его традиционном понимании. В зависимости от того, выполнялся ли в реконструируемом фрагменте речи перевод, конкорданс может быть как обычным, так и параллельным. В некоторых случаях возможна и комбинация обоих типов. Для дискурсивно-переводческого корпуса принципиально важно также то, что в запрос, кроме собственно поискового слова, может включаться различная информация по просодике, коммуникативной функции реплик и т.п. Более того, возможны запросы на получение информации типа «Реплики, в которых председательствующий говорит тихо» или «Реплики, в которых переводчик перебивает подсудимого» и т.п.

Пользователь корпуса получает довольно богатые возможности по сбору самой разнообразной статистики, вплоть до темпа речи говорящих. Так, можно сравнивать темп речи и количество пауз разных переводчиков, а также сравнивать темп речи говорящего и переводчика. Программная оболочка постоянно развивается и пополняется новыми функциями и возможностями.

3. Общие принципы транскрибирования и аннотирования

Для транскрибирования аудиозаписей корпуса была выбрана транскрипция на базе стандартной орфографии на обоих языках (финском и русском) без передачи фонетических особенностей речи говорящих. Причина состоит в том, что фонетическая транскрипция заметно усложнила бы и без того трудоемкую работу, не только замедлив собственно процесс ввода данных, но и сделав невозможной автоматическую лемматизацию. Поэтому от фонетической транскрипции было решено отказаться, поскольку на данный момент использовать материалы для исследований в области фонетики и фонологии не предполагается.

С другой стороны, в транскрипции важно выделить те особенности звучащей речи и ситуации общения, которые потребуются при исследовании дискурса и перевода. Так, для дискурсивного анализа необходимо представление о социальной иерархии участников коммуникативной ситуации. Доминирует в зале суда председательствующий, тогда как между сторонами, т.е. обвинителем и защитником ярко выражено

[2] Авторы выражают признательность лицам, оказавшим помощь в сборе материала и разрешившим получить аудиозаписи из архивов судов, а также их использование в исследовательских целях.

[3] Для обработки текстов используются лемматизаторы RusMorph для русского языка (Гельбух А. Ф., Сидоров Г. О. 2005) и Fintwol для финского.

[4] Рассматривается возможность размещения на сервере аудиозаписей, однако на данный момент это технически неосуществимо в связи с трудо- и времяемкостью работы: необходимо не только «порезать» звуковой файл на реплики, но и убрать из него имена собственные и прочие данные, представляющие конфиденциальный характер.

058-Mihailov_M:_Layout 1 18.05.2008 15:53 Page 378

соперничество, а подсудимый находится на низшей ступени этой иерархии. Речь свидетеля, в свою очередь, изначально должна характеризоваться нейтральностью. Переводчик выступает в роли независимого профессионала, передающего высказывания этих участников коммуникации на другом языке. Мы можем предположить, что хороший перевод отразит языковые личности и роли участников, например, авторитетный характер высказываний судьи. Поэтому указание говорящего оказывается чрезвычайно важным. Кроме того, текстовая передача дискурса без указания на автора реплики вообще вряд ли возможна.

Для дискурсивного анализа большое значение имеет определение структуры дискурса, выявление дискурсивных пар. Судебный допрос строится, как правило, в форме диалога, когда за вопросом следует ответ, а при судебном переводе, соответственно, ВОПРОС → ПЕРЕВОД → ОТВЕТ → ПЕРЕВОД. Однако, на практике такая идеальная конструкция часто нарушается. Причины нарушения дискурсивных пар могут быть различные, невольные или преднамеренные, как например, нежелание подсудимого отвечать на вопрос. Нарушения дискурсивной пары вопрос-ответ могут также происходить из-за неудачного перевода. В корпусе размечены следующие функции высказываний:

Q	вопрос
A	ответ
Qb	подготовка и обоснование вопроса
Foc	уточнение вопроса
Re	повторение вопроса
St	продолжение вопроса
C	приказ/команда
P	предложение
Com	комментирование предыдущего высказывания
Com/Tr	комментирование и/или исправление перевода

и т.д.

Анализ дискурса требует фиксирования многих интонационных и просодических особенностей устной речи, однако, разработка системы транскрипции подчинена задачам и целям исследования (см., напр., Tainio 1997). Цель нашего исследования требует отражения тех просодических сторон устной речи, которые могут быть каким-либо образом связаны с процессом коммуникации. Так, например, темп речи говорящего и хезитация могут быть маркерами уверенности или неуверенности говорящего в достоверности сведений, а интенсивность речи (громкость), фразовая интонация и акцентное выделение могут быть показателем норм коммуникативной ситуации, иерархии участников, разрешения или обязательности выполнения устного волеизъявления.

Нередко то, **как** сказано, может быть важнее того, **что** сказано. Поэтому судебный переводчик в идеале должен отражать все особенности речи говорящего (Driesen 1992; González, Vásquez & Mikkelson 1991: 16, 272; Hale 2004: 8–9; Moeketsi 1999: 100. С другой стороны, у речи переводчика тоже могут быть особенности. Так, например, хезитация или замедленный темп речи могут быть показателем неуверенности переводчика в выборе правильного варианта перевода. При разметке корпуса применяются следующие элементы:

<up>/<down>	движение тона вверх/вниз
<quiet>/<loud>	тихий /громкий
<quick>/<slow>	быстрый / медленный
(.)	пауза 0,2 секунды или короче
(2.9)	пауза длиннее чем 0,2 секунды и ее длина
(ee ~1.0)	пауза, заполненная гласным звуком, и ее длина
(mm ~1.0)	пауза, заполненная согласным звуком, и ее длина

Элементы разметки при выводе на экран заменяются на шрифтовое оформление (например, разрядка для замедленного темпа речи, мелкий шрифт для тихой речи и т.п.) или графические обозначения (например, стрелки для указания движения тона). Приведем небольшой пример (1), чтобы показать, как выглядит транскрибированный диалог при выводе на экран веб-браузера результатов поиска.

(1)

Обвинитель:	**Переводчик**
(11.3) oliko tuota teillä puhetta siellä kun tapasitte Pp:ssa niin (.) tätä enemmänkin kokaiinista?	(0.8) (ee ~0.5) у вас, когда вы встретились в Pp, (0.2) были какие-то (ee) более подробные разговоры о кокаине?
(1.1) {@@@} täällä esitutkinnassa (0.4) kertoneet että, (0.7) kaveri puhui kokaiinista, (.) ehkä sokeri oli kokaiinia.	(0.7) (ee) вы на предварительном следствии показали что, (.) ваш приятель говорил *о* кокаине, и вот этот вот (ee ~0.4) сахар вероятно (ee) был кокаином.
(4.1) ↑ no puhuko ↓ hän jostakin sokerista siinä sitten?	= но ↑ он говорил ↓ вам тогда о чем-то о сахаре?

058-Mihailov_M:_Layout 1 18.05.2008 15:53 Page 379

Обвинитель:	**Переводчик**
(1.7) ↑ kun olette tässä sanonut ↓, että (1.1) että kaveri käski laittaa ↑ repussa olleen sokerin jääkaappiin ↓, tein kuten käski ↑ hän käski ↓, (0.6) kaverini puhui kokaiinista, ehkä sokeri oli kokaiinia.	(0.3) (ее ~0.4) вы (ее) сказали на предварительном следствии, что ваш (.) приятель сказал (0.3) положить (0.3) (ее ~0.3) (.) сахар (0.6) в (.) морозилку, (0.4) и (ее ~0.5) вероятно (0.3) сахар и был кокаином.
(1.0) oletteko kuitenkin (.) siinä vaiheessa ymmärtäneet suomenkielen sanat (.) sokeri (.) jääkaappi (.) kokaiini?	(0.6) вы тогда ↑ все-таки понимали ↓ слова (.) сахар (.) холодильник и (.) кокаин?

4. Вопрос о единице устного перевода

Поскольку корпус предназначен для изучения в первую очередь перевода, основным структурным элементом корпуса должны быть сопоставимые между собой отрезки исходного сообщения и его перевода. Такими **базовыми единицами** в CoInCoUT являются **ИРО** (исходный речевой отрезок) и **ПРО** (перевод речевого отрезка). ИРО – это переводимый сегмент дискурса, т.е. отрезок исходного высказывания/реплики, ограниченный с двух сторон переводом или репликами других участников дискурса. Данный сегмент следует отличать от применяемой в переводоведении «единицы перевода«, которая не имеет четкого формального выражения и может быть как отдельным словом, высказыванием, так и текстом в целом (см., напр., Витренко 2006).

Исходные речевые отрезки (ИРО) и их переводы объединяются в более крупные структуры – **реплики участников** (РУ). Минимальная РУ = ИРО + ПРО, т.е. один исходный речевой отрезок и его перевод. Однако так происходит только в идеале. При реальном судебном допросе используются вопросы различных типов (см. Hale 2004), которые подразумевают ответы разной длины и структуры. Так, например, вопрос типа *Вы были там в 12 часов?* предполагает односложный ответ *Да/Нет.* В свою очередь, на вопрос типа *Расскажите своими словами, как Вы провели этот день?* ожидается достаточно длинный ответ в форме монолога. В РУ такого типа может быть несколько десятков ИРО и ПРО.

Кроме того, в зале суда нередко кипят страсти, и участники коммуникативной ситуации могут перебивать друг друга или переводчика, уточнять или пояснять свое же предыдущее высказывание. В таком случае может происходить нарушение пар ИРО → ПРО. Для разметки наложения реплик участников коммуникации применяется следующая нотация:

[ ] наложение реплик
[/ наложение реплик происходит в середине слова
/] наложение реплик заканчивается в середине слова
= реплика начинается непосредственно после предшествующей без какой-либо паузы.

Далеко не всегда дела обстоят так благополучно, как в примере (2), где обвинитель говорит достаточно медленно и переводчику удается передать подсудимому содержание сообщения. Случается, что русскоязычный участник дискурса, владеющий финским языком, начинает исправлять переводчика, не дослушав перевода до конца. Так, в примере (2) обвинитель спрашивает у свидетеля о событиях, связанных со спором о воспитании ребенка (*lapsen huoltajuuskiista),* переводчик начинает переводить и использует русский термин *спор о воспитании ребенка*, соответствующий употребленному обвинителем финскому термину. Свидетель, не дослушав перевода, перебивает переводчика, говоря о том, что вопрос шел не о воспитании ребенка, а об алиментах, и рассказывает, что тогда произошло. Таким образом, перевод прерван, вопрос обвинителя переведен лишь частично, и на вопрос обвинителя *Стал ли Хх выяснять, где находятся деньги тогда, когда зашел спор* ***о воспитании ребенка****?* был получен ответ *Хх узнал о деньгах, когда встал вопрос* ***об алиментах****.* Ситуацию можно истолковать двояко: (а) свидетель, возможно, исправляет неверное представление обвинителя о положении дел, или (б) свидетель считает неверным вариант перевода, выбранный переводчиком, и исправляет переводчика.

(2)

Обвинитель: **Вопрос**	(1.5) ymmärsinkö (0.2) oikein että, (0.6) kun (0.6) ↑tuli tämän **lapsen huoltajuuskiista** ↓(3.0) niin (0.8) silloin Xx (0.6) ryhtyi (0.3) selvittämään (1.1) näiden (0.9) rahojen kohtaloa, (0.3) missä ne ovat, (0.4) oliko tilillä, (0.4) oli {@@@} ?	**Переводчик:**	(0.6) я правильно понял, что когда (.) возникли (0.2) **спор (0.2)** ***о*** **воспитании ребенка** , тогда [Хх стал] …
Свидетель: **Ответ**	[об **алиментах**]! ↑не о **воспитании** ↓об **алиментах** ! тогда (.) ↑встал вопрос ↓, ↑она сама ↓сказала, ↑что у нее на счету восемьдесят ↓, (0.3) ↑у ребенка на счету ↓восемьдесят тысяч.	**Переводчик:**	(0.7) *se* ei ollut (.) **lapsen huoltajuuskiista** , (.) se oli **lapsen elatuskiista** , (0.4) ja Yy itse (ee) on maininnut että on lapsen ↑tilillä on ↓(.) tämmöisiä varoja .

058-Mihailov_M:_Layout 1 18.05.2008 15:53 Page 380

ИРО и ПРО нередко бывают достаточно длинные, состоят из множества просодически и интонационно обособленных единиц (это видно уже из приведенного выше примера (2)). Для упрощения чтения текста мы вынуждены делить ИРО и ПРО на **элементарные дискурсивные единицы (ЭДЕ)**. Таким образом, базовые единица ИРО и ПРО могут разбиваться на элементарные дискурсивные единицы ИРО = n ЭДЕ или ПРО = n ЭДЕ.

4. Выводы

Из вышесказанного ясно, что корпус устного перевода технически нельзя организовать по той же схеме, по которой создается корпус письменных текстов: в этом случае материалом будет неудобно пользоваться, и значительная часть информации окажется недоступной. При работе над корпусом устной речи, и в особенности – над корпусом устного перевода, становится очевидной многоуровневость и многослойность данных, которые необходимо или желательно представить в таком корпусе текстов. Корпус оказывается необычайно разнообразным и в плане разметки, причем лишь часть ее связана с языковой стороной коммуникации.

Мы назвали корпус устного перевода новым типом корпуса текстов потому, что такой корпус является как бы гибридом одноязычного и параллельного корпуса текстов и требует отражения в разметке информации, связанной с ролями участников общения и ситуациями, в которых это общение происходит. При этом хочется отметить, что часть информации, которую мы пытаемся представить в корпусе **CoInCoUT**, может оказаться весьма полезной и в корпусах письменных текстов. Например, аннотирование диалогической части художественных произведений позволило бы искать примеры употреблений каких-либо выражений в прямой речи. Что касается аннотирования драматических произведений, то здесь применяемая нами разметка оказывается еще более полезной.

Список литературы

1. Витренко А.Г. Что же все-таки такое «единица перевода»? // Вопросы филологии. 2006: 2 (23), с. 53 - 61.

2. Гельбух А. Ф., Сидоров Г. О. К вопросу об автоматическом морфологическом анализе флективных языков // Труды международной конференции «Диалог-2005». М., 2005. (http://www.dialog-21.ru/Archive/2005/Gelbukh%20Sidorov/ GelbukhA.htm, 4.4.2008).

3. Кибрик А.А., Подлесская В.И. К созданию корпусов устной русской речи: принципы транскрибирования. // НТИ. сер. 2. 2003, № 10.

4. Красовская О.В. 2006. Судебный диалог как конвенциональная коммуникативная форма. // Вопросы языкознания. 2006, № 5.

5. Driesen Ch. Status und Funktion des Gerichtsdolmetschers/-übersetzers in Deutschland. // Mitteilungsblatt: Österreicher Ubersetzer- und Dolmetscherverband «UNIVERSITAS«. 1992. S. 7 - 13.

6. Hale S. The Discourse of Court Interpreting. Discourse practices of the law, the witness and the interpreter. Amsterdam/Philadelphia: John Benjamins, 2004.

7. FINTWOL: http://www.csc.fi/english/research/software/fintwol (4.04.2008)

8. González R., Vásquez Victoria F., Mikkelson H. Fundamentals of Court Interpretation. Theory, Policy, and Practice. Durham, North Carolina: Carolina Academic Press, 1991.

9. Mikkelson H.. Introduction to Court Interpreting. Manchester, UK: St. Jerome, 2000.

10. Moeketsi R.H. Discourse in a Multilingual and Multicultural Courtroom: A Court Interpreter's Guide. Pretoria: Van Schaik Publishers, 1999.

11. Tainio L. (ред.) Keskustelunanalyysin perusteet. Tampere: Vastapaino, 1997

12. Välikoski T.-R. The Criminal Trial as a Speech Communication Situation. Tampere: Tampere University Press, 2004.

059-miheev:_Layout 1 18.05.2008 15:54 Page 381

ОПЫТ ВЫБОРОЧНОГО ПОДСЧЕТА ПОЭТИЗМОВ В ПРОИЗВЕДЕНИЯХ В.НАБОКОВА И А.ПЛАТОНОВА[1]

AN ATTEMPT TO COUNT POETIC DEVICES IN EXCERPTS FROM DIFFERENT WORKS OF V.NABOKOV AND A.PLATONOV

Михеев М. *(m-miheev@rambler.ru)*
Научно-исследовательский вычислительный центр МГУ им. М.В. Ломоносова

В выборках примерно по сотне страниц из различных произведений Набокова и Платонова подсчитаны формальные поэтические приемы, встречающиеся у обоих авторов. На основании сравнения их можно далее делать выводы об идиостилях, различающих поэтику одного и другого. В самом общем плане их можно противопоставить друг другу – как поэтику языкового изыска и неказистости.

Интересно сравнить между собой наборы поэтических приемов, которыми работает тот или иной писатель – их можно называть, как уже предлагалось (Григорьев, 2003), экспрессемами или просто тропами, «поэтизмами». С одной стороны, Набоков – как пример классического языка литературы, языка даже несколько гипертрофированно художественного, некий образец использования наиболее присущих в этом отношении средств, в первую очередь именно метафор и сравнений. С другой стороны, Платонов – для контраста и уяснения его собственного вклада в область поэтики (некоторые данные – Михеев, 2000). Мной взяты выборочно отрывки и отдельные произведения этих авторов – примерно по два десятка страниц из начала романа «Чевенгур» (1927-1929) и повести «Котлован» (1929-1930), целиком рассказы «Фро» (1936), «Река Потудань» (1937), «Афродита» (1944) и «Возвращение» (1946), а Набоковские – рассказы «Ужас» (1927), «Возвращение Чорба» (1929), «Соглядатай» (1930), «Весна в Фиальте» (1938), «Сестры Вейн» (1951) и 10 начальных глав романа «Лолита» (1954-1965). Всего около 130 страниц из Набокова и столько же Платонова. (В Таблицах 1 и 2 каждое из обсчитанных произведений зашифровано начальными буквами.) В текстах подсчитаны по возможности все случаи встреченных сравнений, метафор, метонимий, плеоназмов, вульгаризмов, выражений высокого и низкого стилей, контаминаций, двусмысленностей, силлепсисов – в соответствии с (Квятковский, 1966). При этом потенциально бесконечный набор поэтических фигур по возможности сводился к минимуму, многие подрубрики понимались огрубленно и объединялись. Хотелось понять, как на уровне этой – более или менее формальной – поэтики различается творчество двух авторов, в некоторой степени антиподов. Конечно, следующим необходимым шагом должно стать еще и содержательное сравнение – тех способов метафорического, метонимического переноса и прочих видов поэтической трансформации действительности, которые здесь были намечены, можно считать, только на пальцах. Но это, как представляется, дело будущего.

Следует также оговориться, что выявление в живых текстах омертвелых поэтических фигур – занятие, остающееся и на сегодня творческим, подчас раздражая своей нарочитой недискретностью. Имею в виду вынужденное своеволие в принятии решений: часто приходится резать как бы «по живому телу» художественного приема. Обсчитанные тексты и их объемы (в страницах) – на Табл. 1, а следующая за ней Табл. 2 представляет полученные данные употребления поэтических приемов в произведениях Набокова и Платонова. Далее – пояснения к ней:

Коэффициент в Табл. 2 это частное от деления общего числа тропов в произведении на число страниц в нем. Оно условно, показывая нам в некотором приближении наполнение того или иного текста приемами формальной поэтики. Еще одна оговорка: бо́льшее внимание я уделяю все-таки приемам Платонова, его тексты оказываются для меня более насыщенными поэтическими приемами, хотя Набоков – в чем определенный парадокс – безусловно гораздо более благодарный писатель по части того, что обычно называют «поэтическим вкусом». На самом деле очевидно: настоящий знаток Набокова (каковым я не являюсь) сможет насчитать в его текстах и гораздо большее число выделенных здесь приемов, и помимо того, должно быть, выявит такие приемы, которых я здесь вообще не углядел, т.е. существенно увеличит их список. Далее поясняю на примерах, какие именно случаи отклоняющихся от нормы языковых употреблений я отношу к данным тропам, и почему.

[1] Первоначально работа была написана в 2003 г. и прочитана как сообщение на конференции по русской литературе в Воронежском университете. Моя глубокая признательность – памяти покойного Владислава Анатольевича Свительского, пригласившего меня туда.

059-miheev:_Layout 1 18.05.2008 15:54 Page 382

Сравнение: у Набокова русский эмигрант в Берлине, отправившийся на туристическую прогулку в группе немцев с ужасом так описывает кампанию, в которую попал: *Шрам, Шульц и другой Шульц, почтовый чиновник и его жена, все они сливались постепенно, срастаясь, образуя одно сборное, мягкое, многорукое существо, от которого некуда было деваться* («Облако, озеро, башня» – далее сокращенно ООБ)[2] – как нечто вроде дракона или спрута. Платонов при описании сгоревшей электростанции употребляет Олицетворение и формально выраженное Сравнение: *От жара из тела двигателя вытекли все его медные части; сошли и окоченели на фундаменте, как потоки слез, подшипники и арматура...* (А).

Метафора: у Набокова церковный собор в маленьком французском городке назван *длинношеим* (ПМ) – потому что он то и дело *вырастал* перед повествователем, занятого поисками писчебумажной лавки. Или глагольная Метафора: *медленно линяла улыбка на губах* (ООБ). У Платонова же Сравнений и Метафор в целом меньше в пять раз! Причем, платоновские Метафоры – как бы снятые, или переосмысленные языковые, понятые часто нарочито буквально, с приземлением и переразложением[3]. Вот одна из редких у него обычных Метафор: *носильщик со шваброй... начал убирать перрон, как палубу корабля, оставшегося на мели* (Ф). Здесь одна Метафора, отталкивающаяся от сходства предметного (перрон – палуба корабля), и другая, уже от сходства ситуаций – героиня рассказа Фрося остается как корабль на мели без своего мужа Федора, уехавшего на стройку на Дальний Восток.

Загадка – либо с предлагаемой тут же отгадкой (отгадкой, требующей обращения к предтексту, анафорического типа, или с отгадкой в следующем далее тексте, т.е. катафорической, причем, иной раз только в самом конце), или же вовсе без отгадки, что иногда приближает этот употребляющийся Набоковым прием – к платоновскому: *Ее дрожащий рот, кривясь от горечи таинственного зелья, с легким придыханием приближался к моему лицу* (Л) – имеется в виду, по-видимому, *яд страсти* (тут прием Загадка соседствует еще и с приемом Перифраз). Сравним с такой платоновской фразой, где героиня *чувствовала, что в ней самой слабеет сердце от легкости воздуха; все* [в мире] *было слишком отчетливо видно, ослепительно и призрачно – он казался поэтому несуществующим* (Ф). Как воздух может казаться несуществующим, представить довольно трудно. И такого рода загадок довольно много у Платонова.

Образ (или Развернутая метафора) у Набокова: *Ночная бабочка металась по потолку, чокаясь со своей тенью* (ООБ). Формально прием Образ отличается от приема Метафора (а тот – от приема Сравнение), на мой взгляд, только большей сложностью и разработанностью предлагаемой автором транспозиции: в Образе она сложнее, чем в Метафоре, в Метафоре сложнее, чем в Сравнении. В приведенном примере зигзаги полета бабочки метафорически уподоблены чокающимся бокалам, как бы отражающимся в зеркале, а подсвеченный снизу потолок с тенями – зеркалу и его отражениям. У Платонова единый Образ, как правило, не выдерживается, а состоит из разрозненных, друг друга не подкрепляющих, а наоборот, будто опровергающих Метафор и Метонимий.

Метонимия: про отца невесты героя рассказа говорится, что это был *немец в добротном фраке, с рыхлой улыбкой на обезьяньем лице* (ВЧ). Второе из выделенных слов – Сравнение или Метафора, а первое – Метонимия: улыбка человека кажется рыхлой, очевидно, по смежности с упитанным телом или лицом этого немца. У Платонова же метонимический перенос значения – один из основных вообще приемов поэтики.

Олицетворение у Набокова: *мы наскоро обменялись жадными ласками, единственным свидетелем коих были оброненные кем-то темные очки* (Л). У Платонова: *Одно только на прощанье порадовало Захара Павловича – из трубы этой хаты вырос наружу подсолнух, – он уже возмужал и склонился на восход солнца зреющей головой* (Ч).

Издевательство – специфически Набоковский прием: я имею в виду пренебрежение, выражаемое рассказчиком по отношению к герою, а иной раз и к читателю (об этом Адамович, 1994, 227; Михеев, 1999, 96)[4].

Аллитерация: *либидобелиберда* (Л) – где имеется в виду психоанализ; *всё Эдгаровый перегар* (Л) – что, по-видимому, относится к писателям, подражающим Эдгару По. У Платонова прием достаточно редок.

Цитация: *маленькая узкоплечая женщина, с пушкинскими ножками* (Л): Цитата тут совмещена еще и с Загадкой: ясно, что имеется в виду пушкинское выражение *две пары стройных женских ног*, но каких именно, решить затруднительно: если в смысле ограниченного контекста, то должны были быть стройные ножки, но по смыслу расширенного (*Едва ль во всей России сыщешь…*) – скорее кривые: в Набоковском контексте неопределенность сохранена. У Платонова: *надо уметь жить и работать с теми людьми, которые есть на свете* (К), где возможна аллюзия с выражением Маяковского *Отечество славлю, которое есть, / Но трижды –*

[2] Тут и ниже слова, диагностирующие прием, выделяю подчеркиванием.

[3] Относительно Сравнений и Метафор как целого, мной учитывались и чисто грамматические сравнения с союзом как, и стершиеся языковые метафоры, такие как время всю пользу съест (К).

[4] Содержательно строке Издевательство в таблице должна была бы соответствовать строка, описывающая такой контрастный прием Платонова, как авторское самоуничижение, или Кеносис (но его трудно иллюстрировать примерами).

059-miheev:_Layout 1 18.05.2008 15:54 Page 383

которое будет (из поэмы «Хорошо!» 1927). Или *лозунг Ленина: «дьявольски трудное дело управлять государством»* (К) – по-видимому, это чуть ироничное переосмысление известного афоризма вождя о том, что каждая кухарка может научиться управлять государством[5].

Неологизм, или, точнее для Набокова – **Экзотизм**, слово взятое по большей части из словаря В.Даля или же из словаря иностранных слов. (Мной вовсе не принимались в расчет особенности набоковского правописания – такие как *свэтер*, *разстановка, мятель*). Им соответствуют у Платонова, по большей части Вульгаризмы и Аграмматизмы. **Вульгаризм** (или Солецизм, неправильность, элементы Просторечия) – преимущественно платоновский прием[6]: *заквок*[нуть] (Ч)*; отдышься* [от *отдышаться*] (К); *А зачем ты стекло у лампы раздавливаешь?* (В); *...где у него рожались дети* (В). **Аграмматизм** (нарушение управления, приписывание избыточной валентности итп.) – также исключительно платоновский прием: *Каждое сердце разное с другим* (А); *Некуда жить, вот и думаешь в голову* (К).

Еще один специфически платоновский прием можно было бы назвать – **Метонимическим выносом вперед определения** – чаще всего прилагательного, с Эллипсисом причастного оборота или целого придаточного предложения, который оно замещает. В «Котловане» о прощании середняков друг с другом перед коллективизацией (они целуются, как перед смертью) сказано: *Многие, прикоснувшись взаимными губами, стояли в таком чувстве некоторое время, чтобы навсегда запомнить новую родню*. По норме языка *губы* не могут считаться *взаимными* (так же как и руки, ладони, хотя и те и другие совершают движения навстречу друг другу и таким образом *взаимо*-действуют). При этом *взаимными* могут быть отношения (235), расспросы (2), упреки (36), угрозы (5), претензии (45), оскорбления (18), конкуренция (3), согласие (92), не/понимание (37/92), обязательства (46), ласки (11), даже пожатия руки (1) и – объятия[7]. Но не: письма (0), не взгляды (0), не вздохи (0), не болтовня (0)! – в скобках частоты соответствующих сочетаний по Национальному корпусу русского языка. Платоновский оборот заменяет собой что-то вроде: <Многие из них целовали друг друга, проявляя чувства взаимной любви и сочувствия>.

Плеоназм – прием по преимуществу Платоновский: *любил следить за прохожими мимо* (К); *Но наступило позднее время ночи, и Настя закрыла уставшие глядеть глаза...* (В); *Невдалеке от станции строился поселок жилищ* (ГЖ).

Иносказание (сюда же я отнес и Двусмысленность) у Набокова: *вот собственно почему я так благодарно длю это пребывание с маленькой Моникой в кисейно-серой келье воспоминания* (Л): имеется в виду затянувшееся воспоминание, но по-видимому также и сама сцена, его вызвавшая. У Платонова: жена Пашкина уходит, решив отдать Жачеву свой продовольственный паек, чтобы тот, наконец, отстал от нее и ее мужа, – *волнуясь всем не терпящим отлагательства невозможным телом* (К): имеются в виду очевидно и какие-то эротические коннотации.

Ругательства (впрочем, сглаженные литературной допустимостью) у Набокова: *в направлении Бермудских или Багамских или Чертовоматерных Островов* (Л). Более част этот прием у Платонова, с его ориентацией на просторечье и простонародье, т.е. людей с «опушек провинциальных городов»: *Ну что ты будешь делать? Он тут с паровозом как с бабой обращается, как со шлюхой какой!* (Ч). **Эвфемизмы** у Набокова более изысканны: *Цинтия повела меня на второй этаж, в холодную спаленку, чтобы показать мне... разворошенную постель, из которой уже было удалено нежное, несуществующее тело, должно быть знакомое Д.* [ее любовнику, ставшему причиной самоубийства] *до последней бархатистой подробности* (СВ). У Платонова же более откровенны и бесхитростны: *Паровоз никакой пылинки не любит: машина, брат, это – барышня... Женщина уж не годится – с лишним отверстием машина не пойдет...* (Ч). У него вообще-то более употребителен как раз обратный прием, Какофемизм, замена обычного обозначения предмета – более грубым: *Дети просыпались рано, они начинали драться друг с другом в темноте, когда петухи еще дремали, а старики просыпались во втором часу и чесали пролежни* (Ч); Пухов *вспоминал свою умершую от преждевременного износа жену* (СЧ), хотя в целом этот троп я отдельно не обсчитывал.

Оксюморон у Набокова: *пыточная скамья моего блаженства* (Л). Для Платонова данный прием еще более характерен: *...все более уединяясь в тесноте своей печали* (К).

Силлепсис (объединение в одной синтаксической фигуре – обычно в сочинительной конструкции с союзом *и* – двух однородных членов, различных по синтаксическим и семантическим свойствам) у Набокова почти не встречается, у Платонова же, наоборот, весьма и весьма употребителен. Вот пример, где он едва ощутим: *Фрося сидела в сумраке, в блаженстве любви и памяти к уехавшему человеку* (Ф) – переживания любви к уехавшему Федору и память о нем (об их любви) приводят Фросю в состояние блаженства. А вот более существенный: *Машинист... кричал со своего высокого пункта осуждение и указание* (Ф) – то есть кричал,

[5] Допускаю и почти уверен, что не выявленных мной цитат в прочитанных текстах Набокова и Платонова гораздо больше.

[6] Он соотносим, пожалуй, с Макаронизмами и Неологизмами у Набокова.

[7] Правда, последние с устаревшими контекстами – из П.А. Вяземского и В.Т. Нарежного.

059-miheev:_Layout 1 18.05.2008 15:54 Page 384

указывая, но тем самым и осуждал[8]. В более яркой степени проявляющийся Силлепсис, с явным отступлением от грамматики, переходит уже в Анаколуф.

Перифраз у Набокова более редок, чем у Платонова: *древесные листья* (Ф) – вместо просто *листья* или *листья деревьев*. *Мать... говорила дочери слова разумного утешения...* (ВП) – вместо «пыталась утешить» или «приводила разумные доводы, чтобы утешить»; *...слабым голосом сомнения дала знать о своей службе пригородная собака* (К) – то есть, видимо, «негромким ворчанием», как бы «сомневаясь, стоит ли подавать голос».

Конкретизация (когда событие описывается с избыточной подробностью, слишком детально относительно принятого способа описания) у Платонова часто соседствует с Плеоназмом: *Вощев отступил за дверь и скрылся за нею, шепча про себя свою грусть* (К) вместо «шепча что-то себе под нос». У Набокова такого не встретилось, впрочем, этот прием трудно отделить от Экзотизма (см. выше). **Обобщение** (когда событие, напротив, описывается излишне обобщенно, как бы лишаясь необходимых, привычных подробностей, при этом приобретая как бы «взгляд сверху») – опять прием, близкий к Плеоназму: *жалобно пели птицы в освещенном воздухе, не торжествуя, а ища пищи в пространстве* (К); *Сарториус поехал на Крестовский рынок, чтобы купить нужное для своего будущего существования* (СМ).

Так же характерны для Платонова **Канцеляризмы** и Советизмы (когда в речь вторгаются выражения из официального языка, как правило, измененные неправильностью или неверным цитированием): *весь местный класс пролетариата выйдет из мелкоимущественного города; жить ради энтузиазма* (К). Неоправданные, неуместные **Научно-технические термины**: *всматривался... в небесные явления облаков и звезд* (А); *сердце... непрерывно срабатывало текущую кровь в жизненное чувство* (Ф); [Груняхин] *усилил циркуляцию воздуха и сам починил электромотор* (СМ). А вот примеры выражений намеренно **Возвышенного стиля**: *началось деяние его жизни; он приникнет к народу, родившему его* (А); *мать утратила мертвыми всех своих детей* (ВП); *не оставит меня в вечной памяти своей* (Ч). Подобный прием встречается у Набокова, но не в таком масштабе. Сдвиг в отношении нормальной, ожидаемой **Причинности** (а именно, как правило, его гипертрофия – когда логические связи устанавливаются даже там, где их как будто и не может быть): *Из неизвестного места подул ветер, чтобы люди не задохнулись...* (К); *Трава опять отросла по набитым грунтовым дорогам гражданской войны, потому что война прекратилась* (РП). Конструкция с необычным или двусмысленным **Генитивным оборотом**: *в томлении своей разлуки с чужой матерью* (ВП); *общая грусть жизни и тоска тщетности* (К) – (Михеев, 2003, 89-101, 207-236).

Контаминация у Набокова также присутствует, но выглядит скорее как изящный каламбур: *телефонные звонки дальнего следования* (ВФ). А вот у Платонова: *Паровоз курьерского поезда, удалившись, запел в открытом пространстве на расставание* (Ф) – то есть как бы <запел на прощание> и вместе с тем, видимо, <удалившись на достаточное расстояние>.

Вообще **Сложные случаи** наиболее характерны для текстов Платонова – в этом еще одна проблема при подсчете встретившихся поэтизмов – перед нами примеры, как правило, совмещающие в себе сразу несколько тропов: *Назар Фомин стоял тогда возле своих умерших машин* [возле построенной им, но сгоревшей, подожженной кем-то электростанции], *глядевших на него слепыми отверстиями своих выгоревших нежных частей, и плакал* (А) – тут объединены Олицетворение, Эллипсис и Парафраз. Отец Фроси замечает, что его дочь совсем перестала о себе заботиться после отъезда мужа, и говорит ей: *Что ж ты сегодня себе губки во рту не помазала? <...> Иль помада вся вышла? Так я сейчас куплю, сбегаю в аптеку...* (Ф), где, во-первых, разрушено стандартное обозначение действия *мазать губы*, во-вторых, нагнетается избыточность и создается Тавтология *губы во рту* (а где же им быть еще?) и, в-третьих, нарушается нормальное соотношение внутреннего и внешнего, ведь можно *на*красить рот помадой или же *на*нести помаду *на* свои губы, но никак не *помазать губы во рту.* А вот фраза из начала «Епифанских шлюзов», где работающий на строительстве каналов у Петра I английский инженер Вильям Перри пишет своему брату Бертрану на родину: *Четыре года в дикарях живу, и сердце ссохлось, и разум тухнет*. – Здесь Эллипсис, Двусмысленность и Метонимия, которые подразумевают несколько уже существующих в языке Метафор: *живу в дикарях* означает ‘вынужден жить <среди> дикарей’ (по аналогии с: живу *в дальних странах*), но одновременно и ‘сам становлюсь дикарем’ (как: *остаюсь в дураках*). Помимо этого выражение *ссохшееся сердце* тянет за собой *высох*, *ссохся <от любви, тоски, заботы>* или Парономазию *мое сердце <исстрадалось, истосковалось – по родине, по любимой>*, при том что *тухнущий разум* заставляет вспомнить выражения *свет мысли* и *помрачение рассудка*: деятельность разума представляется стандартными метафорами света, огня или костра, освещающего дорогу. Платонов намеренно деформирует, метонимически сдвигая их семантику.

[8] Благодарю японского коллегу Сусуму Нонака за то, что он обратил мое внимание на этот прием у Платонова и, таким образом, пробудил интерес в целом к изучению платоновских тропов.

059-miheev:_Layout 1 18.05.2008 15:54 Page 385

Список литературы

1. Адамович Г. “Наименее русский из всех русских писателей”. Г.Адамович о Вл.Сирине (Набокове) // Дружба народов. 1994 №6.

2. Григорьев В.П. Слова в контекстах русской поэзии XX века (О «Словаре избранных экспрессем») // Известия АН. Серия литературы и языка. Т. 62. 2003. №5. С. 12-23.

3. Квятковский А.П. Поэтический словарь. М. 1966.

4. Михеев М. Заметки о стиле Сирина: еще раз о не-русскости ранней набоковской прозы // Логос. № 11-12. М., 1999.

5. Михеев М.Ю. Жизни мышья беготня или тоска тщетности? О метафорической конструкции с родительным падежом // Вопросы языкознания. М., 2000 № 2.

6. Михеев М. В мир А. Платонова через его язык. Предположения, факты, истолкования, догадки. М.: МГУ, 2003.

Сокращ.	**Тексты Набокова**	Объем (стр.)	Сокращ.	**Тексты Платонова**	Объем (стр.)
ВЧ	Возвращение Чорба	8	А	Афродита	16
ВФ	Весна в Фиальте	20	В	Возвращение	22
Л	Лолита (…)	32	К	Котлован (…)	20
Со	Соглядатай	48	РП	Река Потудань	28
СВ	Сестры Вейн	17	Ф	Фро	20
У	Ужас	6	Ч	Чевенгур (…)	22
	Всего	131			128

Таблица 1. Сокращения, тексты, объемы

059-miheev:_Layout 1 18.05.2008 15:54 Page 386

Произведение / Поэт.прием	**У**	**ВЧ**	**С**	**ВФ**	**СВ**	**Л**	**Вместе**	**Ч**	**К**	**Ф**	**РП**	**А**	**В**	**Вместе**
Время создания: 19..	27	29	30	38	51	65		29	30	36	37	44	46	
Сравнение	6	18	97	80	33	66	**300**	31	4	6	18	1	5	**65**
Метафора	3	6	10	17	13	16	**65**	3	3	2	1		2	**11**
Образ	2	1	5	5	3	9	**25**	4	4	3	3		2	**16**
Загадка	1		2	18	4	6	**31**	3	8	7	5	1		**24**
Издевательство					3	2	**5**							**0**
Цитация						2	**2**		1		2			**3**
Олицетворение	2	3	3	1		2	**11**	6	4	1	6	3		**20**
Аллитерация				1		2	**3**		1					**1**
Неологизм				9		6	**15**	3	1				1	**5**
Вульгаризм							**0**	2	7	9	13	11	5	**47**
Плеоназм							**0**		14		10			**24**
Иносказ-е/ Двусмысл-ть			1	9	2	19	**31**	17	4	3		4	2	**30**
Эвфемизм					1	1	**2**	1	3	1	2			**7**
Ругательство						1	**1**	9						**9**
Оксюморон						1	**1**	1	2				1	**4**
Метонимия / Побоч.смысл	1	3	1	3		3	**11**			1	21			**22**
Парономазия							**0**				3			**3**
Перифраз							**0**		15		5			**20**
Реч.сокращ. / Эллипсис							**0**				9			**9**
Конкретизац.							**0**		6					**6**
Обобщение							**0**	1	21	1	3			**26**
Аграмматизм / Наруш.управл.							**0**		11		6			**17**
Канцеляризм							**0**	3	12		8			**13**
Научно-тех. стиль							**0**	7	5	2	3	2		**19**
Наруш.причин							**0**	5	6	1	5		2	**19**
Возвыш. стиль							**0**	4	13	2	3	12	2	**36**
Генитивная конструкция						1	**1**	13	57	11	4	13	2	**100**
Контаминац. / Каламбур / Силлепсис / Сочинит-ое сокращение					4		**4**	30	81	23	23	6	6	**169**
Синэстезия							**0**				2			**2**
Всего	15	31	119	143	63	137	**508**	143	283	73	155	53	30	**725**
Коэфф-т	2,5	3,9	2,5	7,2	3,7	4,3	**3,9**	6,5	14,2	3,7	5,5	3,3	1,4	**5,7**

Таблица 2. Сравнение поэтизмов Набокова и Платонова

060-Mihina_S:_Layout 1 18.05.2008 15:55 Page 387

К ВОПРОСУ ОБ УНИВЕРСАЛЬНОСТИ ПРОТИВОПОСТАВЛЕНИЯ ИМЕНИ И ГЛАГОЛА

ON THE UNIVERSALITY OF NOUN-VERB DISTINCTION

Михина С.М. *(sofia_mikhina@mail.ru)*
Российский государственный гуманитарный университет

Доклад посвящен типологии языков с ослабленным противопоставлением имени и глагола. При построении классификации в качестве ведущего признака предлагается признак способности/неспособности сочетать именные и глагольные категории внутри одной составляющей, на основании чего выделяются два класса языков с различными свойствами.

Предварительные замечания

Споры об универсальности противопоставления имени и глагола традиционно ведутся на материале определенного набора языков, прежде всего австронезийских (в особенности филиппинских), мунда, древнекитайского, вакашских и салишских (Северная Америка) и западно-кавказских. В настоящей работе предпринимается попытка выделить два типа языков со слабо выраженным противопоставлением имени и глагола и определить основные свойства каждого из типов.

Слабое различение имени и глагола может осуществляться двумя способами. Первый способ состоит в том, что любой корень может присоединять как глагольные показатели, так и именные, но не те и другие одновременно. Такое наблюдается в языках мунда, тонганском, отчасти в английском.

(1) тонганский (Полинезия)

a.	na‘e	ifi	‘a	Sione
	PST	курить	ABS	N.PR
	Джон курил.			

b.	e	kau	ifi
	DEF	PL	курить
	курильщики		

c.	*e	kau	na‘e	ifi
	DEF	PL	PST	курить
	те, которые курили, или *бывшие курильщики*			

(2) английский
* the mistook

В примерах (1c) и (2) показано, что сочетание показателей глагола и имени внутри составляющей вызывает неграмматичность. Корень *ifi* «курить« может присоединять показатель прошедшего времени *na‘e* (1a) или определенный артикль *e* и числовой показатель *kau* (1b), но эти показатели не могут совмещаться друг с другом (1c). Подобным же образом в примере (2) прошедшее время не сочетается с определенным артиклем.

Второй способ выражается в том, что любой корень может «навешивать« на себя глагольные и именные категории вперемешку, и в итоге получается такая часть речи, показатель которой оказывается внешним (вакашские языки, салишские, филиппинские, адыгейский). В примерах (3a)-(4a), где внешний показатель именной, группа используется как имя, а в (3b)-(4b), где внешний показатель глагольный, - как глагол:

(3) адыгейский (Западный Кавказ)
a. s-jE-L&E-StE-r

060-Mihina_S:_Layout 1 18.05.2008 15:55 Page 388

1SG-POSS-муж-FUT-ABS
мой будущий муж

b. rEsijE-m peterburg [jE-stElica-R]
Россия-ERG Петербург 3SG.POSS-столица-PST
*Столицей России (*букв*. в России) был Петербург.*

(4) тагальский (Филиппины)
a. pinatay ko [ang nagaabogado]
PF.PERF.убивать 1SG.ERG ABS AF.IMPF.адвокат
Я убил того, кто становился адвокатом. [Foley 1998]

b. [magdodoktor] na nga talga ako
AF.IRR.врач EMPH EMPH реально 1SG.ABS
На самом деле я буду врачом.

В языках как I типа, так и II типа глагол может выступать в качестве аргумента без каких-либо маркеров перехода в другую часть речи:

(5) тонганский
Ko e nga'ahi fakakaukau 'eni...
DECL DEF многий думать здесь
Есть много идей...

(6) адыгейский
qe-KWa-IWE-StE-r qe-re-KW
DIR-идти-HBL-FUT-ABS DIR-DYN-идти
Кто сможет прийти, пусть приходит.

Возникает вопрос: ведут ли себя аргументы с глагольным корнем полностью аналогично именам или они где-то дают сбой и проявляют свою «истинную«, глагольную природу? Как это происходит в языках обоих типов?

Общие ограничения на аргументы с глагольным корнем

Данные показывают, что аргументы с глагольным корнем не всегда ведут себя аналогично именам. Есть ряд вещей, которые могут делать имена и не могут делать глагольные аргументы. Другими словами, такие аргументы имеют ограничения в своих функциях по сравнению с именами:

✓референциальный статус аргументов с глагольным корнем должен быть маркирован [Mikhina 2007]

Имена могут употребляться с такими средствами обозначения референциального статуса, которые фонетически не выражены, как, например, нулевой артикль в малагасийском или опущение падежного аффикса в адыгейском. Практически всегда это бывают значения, связанные с неспецифичностью [Givon 1984], см. пример (7) из малагасийского:

(7) малагасийский
nangalatra omby ity olona ity
PST.AF.красть корова DEM человек DEM
Этот человек украл (какую-то) корову.

Для аргументов же с глагольным корнем употребление с нулевым артиклем невозможно:

(8) малагасийский
nangalatra ity olona ity *(ny) nentin'ireo[1] ankizy
PST.AF.красть DEM человек DEM DET PST.PF.нести.DEM.PL ребенок
Этот человек украл то, что принесли те дети.

060-Mihina_S:_Layout 1 18.05.2008 15:55 Page 389

Аргумент с глагольным корнем может принимать только те значения из парадигмы специфичности/неспецифичности, которые маркированы. В частности, выражение неспецифичности при таких аргументах грамматично только в тех языках, где маркированы все члены парадигмы выражения референциального статуса, например, в тагальском и тонганском. Неспецифичные аргументы в тагальском маркируются показателем *ng* при глаголах с агентивным вербальным фокусом (см. [Rackowski 2002] и [Aldridge 2006]):

(9) тагальский

bumili	ang	babae	**ng**	**isda**	
PERF.**AF**.покупать	ABS	женщина	ERG	рыба	

Женщина купила (какую-то) рыбу.
**Женщина купила (конкретную) рыбу*. [Aldridge 2006]

Употребление в этой позиции аргументов с глагольным корнем не вызывает неграмматичности:

(10) тагальский

at	ang	pare	at	siya	ay	naghintay
и	ABS	священник	и	3SG	INV	PERF.**AF**.ждать

ng	**sasabihin**	ng	sundalo
ERG	IMPF.PF.сказать	ERG	солдат

И священник и он ждали (того), что скажет солдат. [Bloomfield 1917:30/13]

(11) тагальский

Hayaan	silang	gumawa	**ng**	**kanilang**	**ikinasisiya**
разрешать	3PL-ERG	**AF**.делать	ERG	3PL.POSS-LK	IMPF.PF.любить

sa halip	na	ipinapagawa	ang	ating	kagustuhan	lamang	
вместо	LK	IMPF.PF.CAUS.делать	ABS	1PL.POSS-LK	PF.нравиться	только	

Разрешайте им делать то, что они любят, вместо того, чтобы заставлять их делать только то, что нравится нам.

В тонганском неспецифичность выражается при помощи неопределенного артикля ha; как видно из примера (12), аргументы с глагольным корнем могут использоваться с этим артиклем:

(12) тонганский

fai	ha	tohi
делать.IMPV	INDEF	писать

Напиши письмо. [Churchward 1953:24]

✓аргументы с глагольным корнем, как правило, не могут присоединять показателей посессивности:

(13) адыгейский

***t:jE**-qe-wESE-ZE-xe-Re-r	re-KWa-ZE-Re-x
1PLPOSS-DIR-проснуться-INCH-PL-PST-ABS	DYN-идти-INCH-PST-PL

букв. *Наши кто проснулся ушли*.

(14) малагасийский

*ny	matory-**tsika**
DEF	PRAES.AF.спать-**1PL.INCL.GEN**

букв. *наш, кто спит*

Неочевидно, с чем может быть связано это ограничение; в качестве предварительной гипотезы можно предположить, что проблемы с опознанием референциального статуса, которые вызывают необходимость его маркирования, по какой-то причине не вполне совместимы с семантикой обладаемого.

Дополнительные ограничения

Кроме того, есть дополнительный блок ограничений, которые накладываются только на языки II типа (т.е. те, которые разрешают «наслоение» глагольных и именных категорий в рамках одной составляющей). Эти огра-

[1] Слияние nentiny *принесли* и *ireo* те происходит с помощью правил сандхи.

060-Mihina_S:_Layout 1 18.05.2008 15:55 Page 390

ничения показывают, что аргументы с глагольным корнем устроены по-разному в языках обоих типов. Там, где такие аргументы могут включать в себя глагольные показатели, помимо именных (т.е. в языках II типа), они представляют собой нулевые релятивные обороты, а в языках I типа – нет. Соответственно, для примера (15)

(15) адыгейский
je-s-tE-StE-r hazEr-ew SEtE-R
3SG-1SG-дать-FUT-ABS готовый-ADV лежать-PST
То, что я должен был дать, лежало наготове.

предлагается следующая структура (в духе [Ntelitheos 2006]):

(16) адыгейский
$[_{DP}$ $[_{CP}$ $e_{NP}[_{IP}$ t je-s-tE-StE-]]-r] hazEr-ew SEtE-R
3SG-1SG-дать-FUT-ABS готовый-ADV лежать-PST
То, что я должен был дать, лежало наготове.

В рамках этой гипотезы, нулевая вершина порождается в некоторой позиции внутри клаузы (там, где порождается соответствующий ненулевой аргумент) и перемещается влево, как при обычной релятивизации с ненулевым аргументом, в позицию Spec-CP. Далее рассмотрим, как существование этой структуры проявляется в дополнительных ограничениях на аргументы с глагольным корнем в языках II типа.

- ✓в языках II типа поведение модификаторов при аргументах с глагольным корнем отличается от того, как они ведут себя при именах

Если мы рассмотрим поведение постмодификаторов, то увидим, что в тонганском (язык I типа) они занимают одинаковую позицию при именах и при аргументах с глагольным корнем (см. примеры (17)-(18), постмодификаторы lelei *хороший* и fo‘ou *новый*):

(17) тонганский
a. ha kau fefine lelei
INDEF PL женщина хороший
хорошие женщины

b. [Ko e fakakaukau lelei] pe ia
DECL DEF думать хороший EMPH 3SG

kapau ‘e ‘ave ho’o pepe ki he toketa...
если FUT взять 2SG.POSS ребенок ALL DEF доктор
Это хорошая мысль – взять Вашего ребенка к врачу...

(18) тонганский
a. ‘a e lao fo‘ou
ABS DEF закон новый
новый закон

b. ‘Oku ‘omi [‘e he kau nofo fonua fo‘ou] ha ngaahi
PRAES привести ERG DEF PL жить страна новый INDEF много

lelei fakapa‘anga mo fakasosiale ki he fonua ni...
хороший денежный и социальный ALL DEF страна этот
Новые жители страны привезли в эту страну много материальных и социальных благ...

В то же время в малагасийском (язык II типа) модификаторы, которые в норме должны находиться в постпозиции по отношению к определяемому (см. (19)), при присоединении к глагольному аргументу оказываются не после него (20), а перед ним (21):

(19) малагасийский
zavatra tsara
вещь хороший
хорошая вещь

060-Mihina_S:_Layout 1 18.05.2008 15:55 Page 391

(20) малагасийский

hitako	ny	nafenin-dRabe	tsara
PST.PF.видеть-1SG.GEN	DEF	PST.PF.прятать-Rabe.GEN	хороший

Я нашел то, что спрятал хороший Рабе.
**Я нашел что-то хорошее, что спрятал Рабе.*

(21) малагасийский

hitako	ny	tsara	nafenin-dRabe
PST.PF.видеть-1SG.GEN	DEF	хороший	PST.PF.прятать-Rabe.GEN

Я нашел что-то хорошее, что спрятал Рабе.

Такое поведение модификаторов в малагасийском указывает на то, что в реальности они присоединяются к нулевой вершине, а не к самой глагольной словоформе, и по отношению именно к ней ставятся в постпозицию:

(22) малагасийский

hitako	ny	e_{NP}	tsara	nafenin-dRabe
PST.PF.видеть-1SG.GEN	DEF		хороший	PST.PF.прятать-Rabe.GEN

Я нашел что-то хорошее, что спрятал Рабе.

в языках II типа аргументы с глагольным корнем не могут быть вершиной релятивного оборота

В языках I типа аргументы с глагольным корнем способны к присоединению релятивного оборота, см. пример (23), где группа e kau ifi *те, которые курят, курильщики* присоединяет релятивный оборот ‘oku ta‘ofi ke taimi *бросают вовремя*:

(23) тонганский

Ko	e	kau	ifi	‘oku	ta’ofi	kei	taimi
DECL	DEF	PL	курить	PRAES	бросить	LOC	время

‘oku	nau	moui	fuoloa...
PRAES	3PL	жить	долгий

Курильщики, которые бросают курить вовремя, живут долго...

Что касается языков II типа, то здесь аналогичная конструкция неграмматична:

(24) малагасийский

*ny	nandositra nihomehy		dia	Rabe	
DET	PST.AF.убежать	PST.AF.смеяться		TOP	N.PR

Тот, кто убежал, который смеялся, это Рабе.

Возможна лишь конструкция с сочинением (25), а не с релятивизацией (24):

(25) малагасийский

ny	nandositra sy	nihomehy	dia	Rabe	
DET	PST.AF.убежать	и PST.AF.смеяться	TOP	N.PR	

Тот, кто убежал и смеялся, это Рабе.

Неграмматичность (24) можно объяснить следующим образом. Опять-таки, предположим наличие в обороте ny nandositra *тот, кто убежал* нулевой вершины:

(26) малагасийский

$[_{DP}$	ny $[_{CP}$ e_{NP} $[_{IP}$ t		nandositra]]]
	DET		PST.AF.убежать

тот, кто убежал

Тогда получится, что в примере (24) два несочиненных релятивных оборота относятся к одной вершине, e_{NP}:

060-Mihina_S:_Layout 1 18.05.2008 15:55 Page 392

(27) малагасийский

*[$_{DP}$	ny	[$_{CP}$ e$_{NP}$[$_{IP}$	nandositra]	[$_{IP}$	nihomehy]]]	dia	Rabe
	DET		PST.AF.убежать		PST.AF.смеяться	TOP	N.PR

Тот, кто убежал, который смеялся, это Рабе.

В подавляющем большинстве языков конструкции с множественной релятивизацией, построенные по принципу примера (27), запрещены. Этим запретом, по-видимому, и объясняется невозможность присоединения релятивных определений к аргументам с глагольным корнем в языках II типа.

Итоговая картина распределения ограничений на аргументы с глагольным корнем представлена в Таблице 1.

	языки I типа (тонганский)	***языки II типа (адыгейский, малагасийский, тагальский)***
общие ограничения	референциальный статус аргумента с глагольным корнем должен быть маркирован	
	присоединение посессивных показателей неграмматично	
дополнительные ограничения	---	постмодификаторы ведут себя как премодификаторы
	---	присоединение релятивных оборотов неграмматично

Таблица 1. Ограничения на аргументы с глагольным корнем.

Заключение

То, что глагол не способен до конца вести себя как имя, даже в наиболее «спорных» с точки зрения частеречного деления языках, говорит в пользу универсальности противопоставления имени и глагола. Общие ограничения показывают, чем глагол в принципе отличается от имени: во-первых, ему необходимо маркирование референциальности, т.е. сам по себе он не указывает ни на какого референта, пока не будет эксплицитно выражено, на какого; во-вторых, по причине, которая нуждается в дальнейшем выяснении, есть проблемы с обозначением его посессора. При этом те аргументы с глагольным корнем, где «намешаны» разные категории, ограничены сильнее, чем те, где поверх глагольного корня только именные категории, и в этих дополнительных ограничениях проявляется наличие механизма релятивизации.

На основании этих признаков выделяются два класса языков. Языки I типа (тонганский) не позволяют глагольным и именным категориям сочетаться в рамках одной группы, а в языках II типа (адыгейский, малагасийский, тагальский) эти категории «наслаиваются» друг на друга. Предположительно, устройство двух типов языков принципиально различно: в языках I типа ослабление противопоставления имени и глагола происходит на уровне корня, поскольку выше корня структура слова и предложения выглядит вполне в рамках среднеевропейского стандарта. Языки II типа, напротив, характеризуются тем, что по мере построения словоформы может происходить практически неограниченное количество переходов из одной части речи в другую. Дальнейшее исследование языков обоих типов, возможно, поможет в выявлении глубинной сущности имени и глагола как таковых.

При работе с «проблемными» в отношении лексических категорий языками исследователю нередко не остается ничего другого, как опираться на семантику слова для того, чтобы определить его частеречную принадлежность (см., например, такие работы, как [Broschart & Dawuda 2000], [Himmelmann {в печати}] и др.). Однако этот признак нельзя признать достаточно надежным, поскольку возможно, что членение действительности в рамках исследуемого языка отличается от того, которое навязывается ученому более привычными для него языками. Кроме того, его затруднительно применять в прикладных областях лингвистики ввиду сложностей с его формализацией. Полученные в ходе настоящего исследования результаты могли бы способствовать решению данной проблемы. Выявленные ограничения (см. Таблицу 1) могут быть использованы как формальные критерии для разграничения имени и глагола в тех языках, где решение вопросов, связанных с частеречным делением, не является самоочевидным. Как нам представляется, информация о существовании некоторых универсальных синтаксических отличий между глагольной и именной категорией является важной при решении множества прикладных задач - обучении языку, автоматическом анализе текста, синтезировании предложений системами искусственного интеллекта и т.д.

060-Mihina_S:_Layout 1 18.05.2008 15:55 Page 393

Список сокращений

abs абсолютив, adv адвербиал, af агентивный вербальный фокус, decl декларативная частица, all аллатив, def определенный артикль, dem демонстратив, det детерминатор, dir директив, dyn префикс динамичности, emph эмфатическая частица, erg эргатив, fut будущее время, gen генитив, hbl хабитуалис, impf имперфект, impv императив, inch инхоатив, indef неопределенный артикль, inv маркер инверсии, irr ирреальное наклонение, loc локатив, perf перфект, pf пациентный вербальный фокус, pl множественное число, poss показатель посессивности, praes настоящее время, pst прошедшее время, top топик

Список литературы

1. Aldridge, E. Against Case Agreement in Tagalog. // Paper presented at AFLA XIII, 2006.
2. Bloomfild, L. Tagalog Texts with Grammatical Analysis. Vol.3 // Urbana, Ill.: University of Illinois, 1917.
3. Broschart, Jurgen & Carmen Dawuda. Beyond Nouns and Verbs: Typological Studies in Lexical Categorisation. 2000.
4. http://www.phil-fak.uni-duesseldorf.de/sfb282/working_papers/BEYNOUNF.pdf
5. Churchward, C.M. Tongan grammar. // London – New York – Toronto: Oxford University Press, 1953.
6. Davidson, M. Studies in Southern Wakashan (Nootkan) grammar. // PhD dissertation, 2002.
7. Foley, W. Symmetrical voice systems and precategoriality in Philippine languages. // Paper presented at the 3rd LFG conference, Brisbane, 1998.
8. Givon, T. Syntax: A Functional Typological Introduction.Vol. I // Amsterdam/Philadelphia: John Benjamins Publishing Company, 1984.
9. Himmelmann, Nikolaus. Lexical categories and voice in Tagalog. // In Musgrave, Simon (ed.), Voice and Grammatical Functions in Austronesian Languages. Stanford: CSLI. В печати.
10. Lander Yu.A., Testelets Ya.G. Nouniness and Specificity: Circassian and Wakashan. // Paper presented at PoS 2006, Amsterdam.
11. Mikhina, S. Being a Noun in a Language without Lexical Categories. // Poster presented at the Conference on Nominalizations Across Languages, Stuttgart, 2007.
12. http://web.uni-frankfurt.de/fb10/rathert/forschung/pdfsnom/mikhina.pdf
13. Ntelitheos, D. The Morphosyntax of Nominalizations: A Case Study. // PhD dissertation, 2006.
14. http://dntelith.bol.ucla.edu/dissertation.html
15. Rackowski, A. The structure of Tagalog: Specificity, voice, and the distribution of arguments. // PhD dissertation, MIT, Cambridge, Mass., 2002.

061-Muravenko:_Layout 1 18.05.2008 16:00 Page 394

О СЛОВАРЕ ИЗМЕНЕНИЯ УПРАВЛЕНИЯ В РУССКОМ ЯЗЫКЕ

ON THE DICTIONARY OF CHANGES IN RUSSIAN LANGUAGE GOVERNMENT

Муравенко Е.В. *(emuravenko@yandex.ru)*
Российский государственный гуманитарный университет

В докладе обосновывается необходимость создания нового словаря — словаря изменения управления в русском языке начала XIX – начала XXI вв., предлагается краткая характеристика такого словаря, разбирается пример словарной статьи глагола *скучать*.

1. Об изменении управления в русском языке начала XIX – начала XXI вв.

Современным русским языком в академической традиции принято называть русский язык со времен А. С. Пушкина, т. е. язык последних двух веков. Примеры из классической литературы используются как иллюстративный материал в академических словарях, академических грамматиках, во многих школьных и вузовских учебниках. Однако, обращаясь к произведениям классиков, мы нередко сталкиваемся не только с устаревшими словами, но и с непривычным для нас, устаревшим управлением привычных слов. Рассмотрим некоторые примеры.

Глагол *предать* в значении ‘подвергнуть действию чего-л.’, который сейчас управляет дательным падежом (*предать казни*, *предать осмеянию*, *предать суду* и т. п.), в XIX веке встречается и с управлением *на что*: ***предать на суд*** *демократической истории* (П. А. Вяземский), ***предадим*** *его* ***на произвол*** *благих* (Ф. И. Тютчев), ***предать*** *ее* ***на*** *все* ***муки*** (Ф. М. Достоевский), ***предать*** *себя* ***на служение*** *добрым людям* (Н. С. Лесков). Глагол *обижаться* в XIX веке мог употребляться с существительным в творительном падеже на месте современного винительного с предлогом *на*: ***обидеться*** *его* ***словами*** (А. И. Герцен), ***обижаться*** *нашими глупыми* ***шалостями*** (Ф. М. Достоевский), ***обидеться*** *его внешнею* ***холодностью*** (Д. В. Григорович), ***обижался*** *всякой* ***невежливостью*** (Д. Н. Мамин-Сибиряк).

В некоторых случаях можно обнаружить определенные закономерности, касающиеся целых групп глаголов, напр. у ряда глаголов, имевших ранее предложное управление, оно сменилось беспредложным, но с тем же падежом (*писать к кому-л.* → *писать кому-л.*, *касаться до чего-л.* → *касаться чего-л.*, *достичь до чего-л.* → *достичь чего-л.*): *Я чистосердечно признался в том Марье Ивановне и решился, однако,* ***писать к батюшке*** *как можно красноречивее, прося родительского благословения* (А. С. Пушкин). *Что* ***касается до меня****, то, признаюсь, известие о прибытии молодой и прекрасной соседки сильно на меня подействовало…*(А. С. Пушкин). *Сосед мой, молодой казак, стройный и красивый, налил мне стакан вина,* ***до которого*** *я не* ***коснулся*** (А. С. Пушкин).

Иной раз мы сталкиваемся с тем, что глаголы, которые управляли беспредложным винительным падежом, т. е. были переходными, сейчас имеют другое управление и являются непереходными. Следы прежней переходности мы можем увидеть в страдательных причастиях, относящихся к непереходным глаголам (*руководимый* от *руководить*, *управляемый* от *управлять*, *польщённый* от *польстить* и т. п.). В текстах XIX в. мы можем найти употребления ныне непереходных глаголов в качестве переходных: *Князь Андрей всегда особенно оживлялся, когда ему приходилось* ***руководить*** *молодого* ***человека*** *и помогать ему в светском успехе* (Л. Н. Толстой); *Вы будете* ***руководить юношество*** *по истинному пути* (И. И. Лажечников); *«…завоевания не* ***льстят меня****, но стою за честь»* (Н. М. Карамзин); *Последний, вместе с митрополитом Тукальским,* ***польстил Брюховецкого*** *надеждою, что он останется гетманом, если станет действовать с ними заодно и отступится от Москвы* (Н. И. Костомаров).

К случаям изменения глагольного управления примыкают и случаи изменения управления прилагательных и отвлеченных существительных. Например, существительное *уверенность*, управляющее сейчас только предложным падежом с предлогом *в*, в XIX веке встречается и с формой винительного падежа с тем же предлогом: ***уверенность в*** *наше* ***могущество****, силу и способности* (Д. В. Григорович), ***уверенность в неизменность*** *вашу во всех благородных чувствованиях* (А. Ф. Писемский), ***уверенность в себя,*** *в* ***волю*** *человеческую* (А. И. Герцен), ***уверенность в святость*** (Ф. М. Достоевский) и т. д.

С течением времени меняется и управление предлогов. Так, предлог *сквозь* в XIX веке управлял не

061-Muravenko:_Layout 1 18.05.2008 16:00 Page 395

только винительным, но и родительным падежом: *И долго, будто* ***сквозь тумана****, Она глядела им вослед*... (А. С. Пушкин). Сейчас этот предлог употребляется только с винительным падежом. Предлог *между*, использовавшийся как с родительным, так и с творительным, в настоящее время управляет творительным; употребление родительного воспринимается как архаичное: *Всюду между деревьев ... мелькали белые, синие, красные рубахи* (И. С. Тургенев). Существенно изменилась картина управления предлога *по*: управление предложным падежом во многих значениях постепенно сменилось на управление дательным, ср. употребления в XIX в.: *Два раза они замечали, что внизу, близко от них, показывались французы, и тогда они* ***били по них картечью*** (Л. Н. Толстой); *На лице виднелось удивление, вопрос, и еще какое-то особенное возбуждение* ***проходило по нем*** *быстрыми тенями* (В. Г. Короленко); *Каждую минуту* ***тосковать по прошлом****, следить за успехами других, бояться смерти... не могу!* (А. П. Чехов). (Об изменении управления предлога *по* см. подробно в [Муравенко 2006].)

2. О словаре изменения управления

В школьных изданиях классической литературы, где принято комментировать историзмы и лексические архаизмы, приведенные выше случаи устаревшего управления обычно никак не поясняются. В словарях эти случаи отражаются весьма непоследовательно. Мне не раз доводилось слышать, как школьники расценивали их как ошибки писателей. Сложившаяся ситуация говорит о насущности создания словаря нового типа — словаря изменения управления в русском языке XIX—XXI вв.

2.1. Место словаря изменения управления в типологии словарей

2.1.1. Словари синхронные vs диахронические, современные vs исторические

Определяя место предлагаемого типа словаря в общей типологии словарей, я опираюсь на классификацию словарей, содержащуюся в [Шимчук 2003], — наиболее полную и современную из известных мне классификаций.

Лингвистические словари автор делит на синхронные и диахронические. «Это противопоставление задается ориентацией на отражение в рамках словаря единой системы в определенный момент времени («синхронный» подход) или на упорядочивание диахронических языковых данных по хронологическому принципу, а также на реконструкцию изменений, пережитых единицами этой системы с момента их появления. К диахроническим словарям относятся исторические, в которых лексика языка представляется во временных изменениях (обычно — в пределах заданного периода), и этимологические, содержащие информацию о происхождении слов» [Шимчук 2003: 13].

Такое разбиение кажется вполне логичным, однако, судя по приведенной в книге типологии словарей, данные выше определения понятий «синхронный» и «диахронический» не вполне соответствуют реальному использованию этих терминов автором при создании классификации. Так, словари древних памятников и «Словарь русского языка XVIII века» относятся к историческим, хотя охватывают небольшой период времени, словари же современного русского языка, охватывающие более продолжительные временные отрезки, называются синхронными. «Словарь Академии Российской» (1789–1794) разбирается среди синхронных, а «Словарь русского языка XVIII века», издаваемый в настоящее время, — среди диахронических (точнее, исторических).

Таким образом, к историческим словарям как в [Шимчук 2003], так и в других лексикографических трудах относятся словари русского языка, в которых метаязык описания не совпадает с описываемым языком (это словари древнерусского языка и русского языка до «пушкинской поры»), а среди синхронных Э. Г. Шимчук перечисляет словари, в которых метаязык описания совпадает с описываемым языком, т. е. словари русского языка, издаваемые с конца XVIII в. и ориентированные на представление лексической системы языка, современной авторам словарей.

Правда, как замечает автор классификации, «противопоставление исторический (диахронический) — современный (синхронный) словарь … не носит абсолютного характера» [Шимчук 2003: 154], «элементы историзма могут быть и в синхронном словаре современного языка» [Шимчук 2003: 14]. Как «синхронный словарь с элементами историзма» квалифицируется Большой ***Словарь современного русского языка*** (БАС) [Шимчук 2003: 20].

Для ясности дальнейшего изложения я должна остановиться на понятии «современный русский язык», которое в русистике трактуется по-разному. Одни лингвисты считают, что современный русский язык — это язык со времен А. С. Пушкина до наших дней (см. выше). Словари, отражающие современный русский язык в таком понимании (далеко не только БАС) не являются, строго говоря, синхронными. Другие лингвисты под

061-Muravenko:_Layout 1 18.05.2008 16:00 Page 396

современным русским языком понимают язык с середины XX в., т. е. язык, активно использующийся в настоящее время его носителями.

Мне представляется, что, классифицируя словари современного русского языка, к синхронным в строгом смысле слова правильно было бы относить только словари современного русского языка во втором понимании, т. е. словари, в которых отражается активная лексика. Словари же, описывающие язык двух столетий, в принципе не могут быть синхронными, поскольку за такое время язык существенно меняется.

Лингвистические описания и, соответственно, словари, которые отражают синхронное состояние языка с учетом происходящих в нем динамических процессов, можно назвать «бисинхронными» (этот термин использован вслед за Полем Гардом в [Кронгауз 2006: 3]).

Именно бисинхронным я бы считала словарь изменения управления в русском языке последних двух веков, т. е. в языке, называемом в академической традиции современным русским языком.

2.1.2. Словари неологизмов и устаревших слов

Среди аспектных словарей принято выделять, с одной стороны, словари новых слов и словари языковых изменений, а с другой, словари устаревших слов.

Казалось бы, именно в этих двух группах словарей можно найти словари, отражающие изменения синтаксических свойств слов. Однако под языковыми изменениями авторы понимают лишь появление новой лексики и новых значений у слов (см., например, Толковый словарь современного русского языка. Языковые изменения и норма 20 века / Под ред. Г. Н. Скляревской. М., 2001). В словарях устаревших слов рассматриваются историзмы и лексические архаизмы. Слов же с изменившимися синтаксическими свойствами, т. е. синтаксических архаизмов, мы в этих словарях не найдем.

2.1.3. Словари управления

Управление — важная характеристика прежде всего глаголов, а также и многих других лексем. Информация об управлении слова должна содержаться в хороших толковых словарях, однако, как справедливо замечено в [Шимчук 2003: 27], «синтаксическая информация о слове в большей части традиционных словарей носит эпизодический характер». Информацию об управлении и сочетаемости соответствующей единицы предполагается последовательно давать в новом «Интегральном словаре русского языка» (см. [Апресян 1986]), работа над которым пока не завершена.

Существуют специальные словари управления в русском языке. Их не так много, и в них отражается определенный срез языка. К основным словарям управления можно отнести следующие: [Муцков, Влахов 1966], [Прокопович и др. 1975], [Денисов, Морковкин 1978], [Розенталь 1981], [Апресян, Палл 1982] (о принципах составления словаря управления см. в [Апресян 1996]).

Перечисленные словари относятся к синхронным.

Словарей изменения управления пока не существует.

2.2. Краткая характеристика словаря изменения управления

В качестве материала для задуманного словаря будут отобраны те слова (глаголы, прилагательные, существительные и предлоги), управление которых менялось на протяжении двух последних столетий, а также те, которые имеют вариативное управление в настоящее время. Предполагается 1) показать современное (актуальное) управление слова, что необходимо для его правильного употребления; 2) описать все возможные способы управления на протяжении двух веков с указанием хронологических границ каждого способа, что не только познавательно (поскольку проясняет и появление современного способа управления), но и необходимо для правильного понимания классической литературы. Словарь может быть полезен и для целей автоматической обработки текстов, в частности машинного перевода: расширенное описание управления слов потенциально увеличивает полноту синтаксического анализатора, а указание хронологических границ каждого способа управления дает возможность выбрать из всех вариантов управления наиболее адекватный при синтезе текста.

Толкование в общем случае не дается, а указывается тогда, когда разные способы управления связаны с разными значениями, а также в случае происходящего со временем сужения или расширения значения, не сопровождающегося изменением управления.

Пока не решен вопрос о границах понимания термина «управление». Указывать ли в словаре только те формы, которые являются выражением переменных в его толковании, т. е. заполняют валентности, или рассматривать более широкую сочетаемость слова? Этот вопрос можно будет решить только по мере накопления материала.

Ясно, что описание управления каждого слова может быть адекватным только с учетом системных связей в языке. Об этом подробнее см. при описании глагола *скучать*.

В качестве опорных источников предполагается использовать все толковые словари и словари

061-Muravenko:_Layout 1 18.05.2008 16:00 Page 397

управления, изданные в XIX и XX вв., в качестве источника употреблений — прежде всего компьютерные базы текстов и в первую очередь Национальный корпус русского языка. Однако, используя изданные в XX в. произведения XIX в., необходимо отдавать себе отчет в возможности поздних редакторских исправлений, а пользуясь компьютерными базами — в возможности ошибок сканирования и других сбоев. Поэтому в особых случаях необходима тщательная проверка примеров.

2.3. Изменение управления глагола «СКУЧАТЬ»

Возьмем в качестве примера глагол *скучать*. У этого глагола за два века произошли существенные изменения в управлении, что связано, во-первых, с изменением значения глагола, во-вторых, с выравниванием его управления по аналогии с другими глаголами близкой семантики, в-третьих, с уменьшением в языке числа инфинитивных дополнений, в-четвертых, с изменением управления предлога *по* и, наконец, с изменением способов выражения причинного значения. Как видим, причины изменений в управлении носят системный характер, а не касаются только данного глагола.

Рассмотрим все значения и возможные способы управления этого глагола в хронологическом порядке. В XVIII веке глагол *скучать* употреблялся в значении ‘надоедать, вызывать скуку’ и имел управление «что скучает кому» и «кто (что) скучает кому чем» (диатеза с расщеплением) — ср. с современными глаголами *докучать*, *наскучить*. В начале XIX века это значение встречалось уже редко и постепенно совсем исчезло из языка (исчезло и соответствующее управление). Последний по времени пример, встретившийся в НКРЯ с этим значением, относится к 1858 году: *«Я собой никому не скучаю, прошу и мне не* скучать*»*. (С. Т. Аксаков).

Другое значение — ‘испытывать скуку’. Как любое эмоциональное состояние, состояние скуки обычно бывает вызвано какими-то причинами. Для обозначения причины прежде всего использовались стандартные способы: *чем* (вплоть до конца XIX в.), *за чем-л.* — до конца XIX века, *от чего-л.*, *из-за чего-л.* — с середины XIX века, придаточные предложения). Примеры: *Во время этой второй болезни,* ***скучая долгою праздностью****, он просил книг: ему дали Flores Sanctorum* (Т. Н. Грановский. 1849–1850). *Однажды, во время пребывания двора в Гатчине, генерал-прокурор (Петр Хрисанфович Обольянинов), воротясь от императора с докладом, объявил Безаку, что государь* ***скучает, за невозможностью*** *маневрировать в дурную осеннюю погоду, и желал бы иметь какое-либо занятие по делам гражданским* (Н. И. Греч. 1849–1856). *Иван Петрович, обратив внимание, что Егор* ***скучает от бездействия****, принес ему пилок и дощечек от сигарных ящиков с наклеенными узорами и научил его подставочки выпиливать* (Н. С. Лесков. 1880).

В начале XIX в. (редко — позже) глагол *скучать* мог управлять инфинитивом:
Счастлива еще она была, что нрав ее и разум ставили выше всех семейных раздоров и что она ***не скучала сидеть одна дома*** *с своей работой или книгами и выезжать не любила, впрочем же ни с кем не вела знакомства*. (И. М. Долгоруков. 1799-1806). Пример из БАС: *Они* ***не скучали ухаживать*** *за мною в эту долгую зиму* (Н. Г. Чернышевский). В современном языке инфинитив с глаголом *скучать* уже не употребляется, что является отражением общей истории развития инфинитивного дополнения.

Помимо перечисленных стандартных способов выражения причинной валентности, глагол *скучать* допускает и другие формы для выражения этого значения, которые можно назвать косвенными способами. Это придаточные предложения с союзами *если* и *когда*, деепричастные обороты, предложно-падежные группы и наречия с пространственным значением (*в деревне*, *на новом месте*, *за прилавком*), предложно-падежные группы, имеющие значение ‘во время какого-л. действия’ (*за вышиванием*, *за уроками* ‘занимаясь уроками’, *над книгой* ‘во время чтения книги’, *на лекции*), группа *с кем* со значением совместности. Все эти косвенные способы заполнения валентности используются на протяжении всего рассматриваемого периода времени, т. е. с начала XIX в. по настоящее время, однако в современном языке встречаются чаще.

В значении ‘испытывать скуку’ глагол может употребляться и без дополнения, однако в этом случае, как правило, мы сталкиваемся с контекстным выражением валентности, например: *Послы сидели одни, не могли заводить знакомств и скучали* (Н. М. Карамзин. 1813–1820.

Если состояние скуки было вызвано отсутствием кого-чего-л., использовалась (и в настоящее время также используется) форма «*без кого-чего-л.*»: *Я узнал, что он в последние годы совсем стал хилый, почти в детство впал, так что даже* ***скучал без игрушек*** (И. С. Тургенев. 1874); *Я чувствовал, что моя дочь скучает* ***без меня*** (Ю. Визбор. 1983). Постепенно значение ‘скучать без кого-чего-л.’ становится основным, в конце первой трети XIX века в этом значении уже возможно употребление глагола с управлением «*о* + предл.п.»: *После этого меня отвезли обратно к бабке; я долго* ***скучала о молодом Кирияке*** (Н. А. Дурова. 1835). Примерно тогда же появляется управление с предлогом *по* под влиянием управления глаголов *грустить*, *тосковать*. Предлог *по* в начале XIX века управляет в основном предложным падежом, но возможен и дательный. Предложный постепенно вытесняется дательным. Временна́я последовательность замен такова: существительные во

061-Muravenko:_Layout 1 18.05.2008 16:00 Page 398

множественном числе — местоимение *они* — существительные в единственном числе — местоимения *он, кто* — местоимения *мы, вы*. Поскольку существительные в форме предл. падежа. мн. числа после предлога *по* перестали употребляться еще до того, как глагол *скучать* стал управлять сочетаниями с предлогом *по*, то конструкции *скучать по* + сущ. в предл.мн. ни в какой момент времени не существовали, а все остальные возможности были представлены в языке и зафиксированы в текстах. Что касается местоимений *я*, *ты* и *она*, то омонимия форм дательного и предложного этих местоимений сделала переход более гладким и незаметным. Примеры с предл. падежом: *По ночам он спал теперь нехорошо, чутко, и ему слышно было, как Матвей тоже не спал и всё вздыхал,* ***скучая по своем изразцовом заводе*** (А. П. Чехов 1895). *Может быть, она все-таки сейчас* ***скучает по нем***? (В. В. Набоков. 1927–1928). *Этим я хочу сказать, что* ***скучаю по Вас*** (Б. Ефимов. 2000). *И* ***ни по ком скучать не будете***? (Л. Юзефович. 2001). Примеры с дат. падежом: *Кто из нас не встречал, например, работниц и кормилиц в городах,* ***скучающих по пашне и сенокосу***? (Ф. М. Достоевский. 1880). *— Ну, нам час добрый, а вам счастливо оставаться,* ***по нам не скучать*** (И. Шмелев. 1930–1931). *Я* ***скучаю по Вам****, лаборатории и в особенности по моей научной работе* (П. Капица. 1998).

Попробуем упорядочить изложенную информацию, полученную в результате анализа данных НКРЯ и некоторых произведений XIX в., не вошедших в НКРЯ, в словарной статье глагола. Для составления статьи учитывались также данные перечисленных выше в пункте 2.1.3. синхронных словарей управления и толковых словарей БАС, МАС, СУ, СО, СЯзП. Предлагаемый вариант словарный статьи носит предварительный характер и нуждается в тщательной проверке, привлечении дополнительных данных и редактировании (дополнительные источники особенно необходимы для более точной датировки употребления форм 1 и 2 в значении II, поскольку в НКРЯ и других проанализированных источниках для этого недостаточно примеров). Предполагается, что после словарной статьи будут приведены иллюстративные примеры всех возможных случаев.

СКУЧАТЬ (начало XIX в. – начало XXI в.)

1. ‘Кто-л. испытывает досаду, недовольство, тяготится чем-л.’ (до начала XX в.)**;**
‘кто-л. испытывает скуку’

1) с инфинитивом (до начала XIX в., редко — позже): *она скучает сидеть дома*;
2) стандартные способы выражения причины —
кто скучает чем (до конца XIX в.): *он скучает праздностию, молчанием, однообразностью, поучениями кого-л.*;
кто скучает за чем (до конца XIX в., редко): *он скучает за невозможностью делать что-л.*;
кто скучает от чего (со 2-й половины XIX в.): *он скучает от безделья, от одиночества*;
кто скучает из-за чего (со 2-й половины XIX в., редко): *она скучала из-за плохой погоды*;
с придаточными причины (редко): *мальчик скучал, потому что не мог привыкнуть к новой школе*;
3) косвенные способы выражения причины **—**
кто скучает, когда…: *мальчик скучал, когда оставался один***;**
кто скучает, если…: *она скучала, если нечего было читать***;**
кто скучает + деепричастный оборот**:** *он скучал, отвечая на ее вопросы*;
кто скучает где: *она скучает в деревне, на новом месте, у бабушки, здесь***.**
кто скучает за чем, над чем (занимаясь чем-л.)**, на каком-л. мероприятии:** *он скучает за посадкой картофеля, над учебниками, на лекции***;**
кто скучает с кем: *он не скучает с ней***.**

2. ‘Кто-л. болезненно переживает отсутствие кого-чего-л.’ (с 30-х гг. XIX в.)

1) **кто скучает о ком-чем** (с конца 30-х гг XIX в., в наст. время редко): *дети скучают о матери, об отце, она скучает о дочери, о сыне, о муже, он скучает о жене, о родных, о семье, о доме*;
2) **кто скучает по ком-чем**
(*по них* — до середины XIX в., редко — позже);
(с сущ. в ед.ч. — до конца XIX в.): *она скучает по муже, по сыне, по доме*;
(*по ком*, *по нем* — до середины XX в., редко — до конца XX в.);
(*по вас*, *по нас* — по наст. время, в XXI в. редко).
3) **кто скучает по кому-чему** (по наст. время)
(с сущ. во мн.ч.): *он скучает по родным, по друзьям*;
(*по ним* — с 30-х гг. XIX в.);
(с сущ. в ед.ч. — с середины XIX в.): *она скучает по мужу, по сыну, по дому*;
(*по кому*, *по нему* — с середины XIX в.);
(*по вам*, *по нам* — с конца XX в.);
4) **кто скучает без кого-чего**: *я скучаю без тебя, без родных, без друзей, без телефона, без игрушек*;
5) **кто скучает за кем-чем** (обл. южн.): *я скучаю за тобой*;
6) косвенные способы выражения участника ситуации **—**

061-Muravenko:_Layout 1 18.05.2008 16:00 Page 399

кто скучает, когда…: *мальчик скучал, когда уходили родители***;**
кто скучает, если…: *она скучала, если его долго не было***;**
кто скучает + деепричастный оборот**:** *он скучал, долго не видя ее*.
3. ‘Что-л. (кто-л.) вызывает скуку; вызывает досаду, недовольство; надоедает’ (до середины XIX в.)
1) **что скучает кому**: *ему скучает бездействие, им скучают лишние издержки, вопросы мальчика скучают отцу*;
2) **кто (что) скучает кому чем**: *мальчик скучает отцу вопросами*.

Список литературы

1. Апресян 1986 — Апресян Ю.Д. Интегральное описание языка и толковый словарь // Вопросы языкознания. 1986. № 2. С. 57–70.

2. Апресян 1996 — Апресян Ю.Д. О толковом словаре управления и сочетаемости русского глагола // Словарь. Грамматика. Текст: Сб. ст. / РАН. Отд. лит-ры и языка. Ин-т рус. яз. им. В. В. Виноградова / Отв. ред. Ю. Н. Караулов, М. В. Ляпон. — М.: Азбуковник, 1996. С. 13–43.

3. Кронгауз 2006 — Кронгауз М.А. От составителя // Изменения в языке и коммуникации: XXI век. М.: Изд-во РГГУ, 2006. С. 1–4.

4. Муравенко 2006 — Муравенко Е.В. О синтаксических архаизмах // Изменения в языке и коммуникации: XXI век. М.: Изд-во РГГУ, 2006. С. 209–224.

5. Шимчук 2003 — Шимчук Э.Г. Русская лексикография: Учебное пособие. М.: Изд-во МГУ, 2003. 320 с.

Словари

1. Апресян, Палл 1982 — Апресян Ю.Д., Палл Эрна. Русский глагол — венгерский глагол. Управление и сочетаемость. Будапешт, 1982, Т. 1–2.

2. Денисов, Морковкин 1978 — Учебный словарь сочетаемости слов русского языка / Под ред. П.Н. Денисова, В.В. Морковкина. М., 1978.

3. Муцков, Влахов 1966 — Муцков Л., Влахов С. Беспредложное и предложное управление глаголов в русском языке (словарь-справочник). София, 1966.

4. Прокопович и др. 1975 — Прокопович Н.Н., Дерибас Л.А., Прокопович Е.Н. Именное и глагольное управление в современном русском языке. М., 1975.

5. Розенталь 1981 — Розенталь Д.Э. Управление в русском языке: Словарь-справочник. М., 1981.

6. БАС — Словарь современного русского литературного языка: в 17-и т. / АН СССР, Ин-т рус. яз.; Под ред. В.И. Чернышева, С.П. Обнорского, В.В. Виноградова, Ф.П. Филина и др. М., 1948–1965.

7. МАС — Словарь русского языка: В 4-х т./ АН СССР, Ин-т рус. яз.; Под ред. А.П. Евгеньевой. — 2-е изд., испр. и доп. — М.: Русский язык, 1981.

8. СО — Ожегов С. И. Словарь русского языка: 70.000 слов / Под ред. Н.Ю. Шведовой. — 22-е изд., стер. М.: Русский язык, 1990.

9. СУ — Толковый словарь русского языка в 4 т. под ред. Д.Н. Ушакова. М.: ОГИЗ, 1935–1940.

10. СЯзП — Словарь языка Пушкина: в 4 т./ Отв. ред. акад. АН СССР В.В. Виноградов. — 2-е изд., доп. М.: Азбуковник, 2000.

Основной источник:

Национальный корпус русского языка — www.ruscorpora.ru

062-Nedoluzhko:_Layout 1 18.05.2008 16:01 Page 400

СИНТАКСИЧЕСКИ АННОТИРОВАННЫЙ КОРПУС ЧЕШСКОГО ЯЗЫКА THE PRAGUE DEPENDENCY TREEBANK

Недолужко А. *(nedoluzko@ufal.mff.cuni.cz),* ***Гаич Я.*** *(hajic@ufal.mff.cuni.cz),* ***и кол.***

Институт формальной и прикладной лингвистики, физико-математический факультет, Карлов университет, Прага, Чехия (ÚFAL MFF UK)

The Prague Dependency Treebank (PDT 2.0) – это корпус текстов чешского языка, аннотированный на трех связанных между собой уровнях – морфологическом (2 млн словоупотреблений), поверхностно-синтаксическом (1.5 млн) и глубинно-синтаксическом (0.8 млн). На глубинно-синтаксическом уровне аннотируется также актуальное членение предложений и именная кореференция. PDT 2.0 основан на пражской лингвистической традиции, адаптированной к требованиям современной компьютерной лингвистики. Аннотация корпуса проводится частично автоматически.

Помимо обширного корпуса чешских текстов разрабатывается проект параллельных текстов на чешском и английском языках (The Prague Czech-English Dependency Treebank), где подобным образом аннотируются тексты из Wall Street Journal и их переводы на чешский язык. Целью проекта является подготовка текстовой базы для обучения компьютера машинному переводу.

В реферате я представлю общую схему аннотации с особым акцентом на глубинно-синтаксический уровень, расскажу о системе синтаксических функторов узлов на этом уровне и словаре моделей управления предикатов, встроенном в проект, а также отвечу на все возникшие вопросы.

1. Общие сведения

Синтаксически аннотированный корпус чешского языка (PDT) – это проект лингвистического (морфологического, синтаксического, семантического, прагматического и др.) аннотирования текстов, разрабатываемый в настоящее время в Институте формальной и прикладной лингвистики физико-математического факультета Карлова университета в Праге. Последняя версия проекта, PDT 2.0, содержит большое количество чешских текстов (2 млн. словоупотреблений) с аннотацией (взаимосвязанной) на трех уровнях – морфологическом (2 млн. слов), поверхностно-синтаксическом (1.5 млн. слов) и глубинно-синтаксическом (0.8 млн. слов). Корпус использует самые современные способы аннотации (раздельная аннотация уровней с использованием XML, RelaxNG). К корпусу также прилагается отдельная поисковая программа Netgraph, позволяющая производить сложный поиск по многим параметрам и собирать материал и статистические данные для лингвистических исследований.

Аннотирование синтаксических уровней производится вручную на основе предварительных автоматических аннотаций, т.е. фактически аннотирующий лингвист просматривает уже готовую аннотацию, дополняет ее и исправляет ошибки. Аннотирование синтаксических уровней проводится с помощью специальной программы для аннотирования корпусных данных TrEd (од *tree editor*), разработанная на ÚFAL MFF UK. Аннотирование вручную проводится аннотаторами с лингвистическим образованием, причем регулярно проводится тест на т.наз. «соответствие аннотаторов», т.е. все аннотаторы, работающие на данном проекте, аннотируют одни и те же тексты, на которых затем проводится автоматическая проверка соответствия.

Лингвистическая основа PDT восходит к традициям пражской лингвистической школы и функционально-грамматическому описанию языка, разработанному в шестидесятых годах двадцатого века чешским лингвистом П. Сгаллом и его учениками.

PDT – один из нескольких десятков проектов синтаксически аннотированных корпусов, разрыбатываемых в настоящее время в мире. Идейным вдохновителем проекта послужил американский PennTreebank (http://www.cis.upenn.edu/~treebank), однако со структурной точки зрения он значительно отличается от PDT и разработан на основе принципа непосредственных составляющих. Лингвистически близким PDT является разработка И.Богуславского и система уровней ЭТАПа-3, но в PDT несравнимо большую роль играет статистика,

062-Nedoluzhko:_Layout 1 18.05.2008 16:01 Page 401

иначе работает система синтаксических отношений, больше объем обработанного автоматически и вручную материала и т.д. С т.з. количества синтаксически обработанного материала PDT можно сравнить с корпусом датских текстов Danish Dependency Treebank – 5500 синтаксически аннотированных деревьев (http://www.id.cbs.dk/~mbk/treebank), португальских текстов – The Floresta Sintá(c)tica project, 10000 деревьев (http://acdc.linguateca.pt/treebank/info_floresta_English.html), турецких текстов – METU-Sabanci Turkish Treebank (http://www.ii.metu.edu.tr/~corpus/treebank.html) и др. Несомненным преимуществом PDT является комбинация большого количества аннотированных текстов с богатой лингвистической информацией, в т.ч. выходящей за рамки одного предложения (аннотация кореференции, актуального членения, сочинительных конструкций и др.)

Автор данного доклада, А. Недолужко, ведет аннотацию кореференции на тектограмматическом уровне, а также занимается синтаксической аннотацией английских предложений для проекта PEDT (см ниже), Ян Гаич является руководителем всего проекта.

2. Уровни аннотации

Аннотирование проводится на трех уровнях – морфологическом, поверхностно-синтаксическом и глубинно-синтаксическом. В действительности существует еще нулевой уровень основного текста, где всем элементам (слова, числа, знаки препинания) присваиваются идентификаторы. На рис. 1 изображена связь между уровнями: так, как они аннотируются в PDT 2.0. Это разбор чешского предложения *Byl by šel dolesa (Шел бы в лес*), содержащее глагол «идти» в сослагательном наклонении в прошедшем времени (*Byl by šel*) и опечатку (*dolesa «в лес»* написано слитно, должно быть *do lesa*).

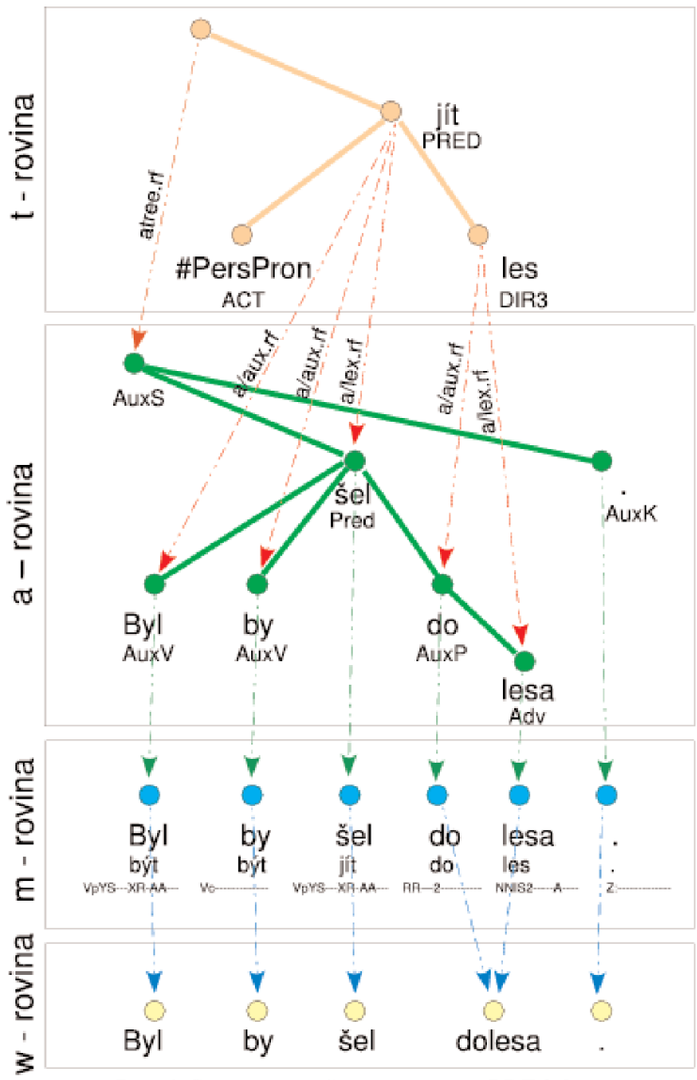


Рис. 1 Связи уровней аннотации в PDT 2.0

062-Nedoluzhko:_Layout 1 18.05.2008 16:01 Page 402

2.1 Морфологический уровень

Здесь словоупотреблениям нулевого уровня присваивается некоторое количество атрибутов, из которых самыми важными являются морфологические: lemma и tag. Атрибут lemma представляет собой имя лексемы данного слова и однозначно соотносит его с морфологическим словарем. Атрибут tag содержит 15 позиций морфологической информации (часть речи и все актуальные для нее морфологические характеристики, напр. NNIS2——-A——). Пример аннотации на морфологическом уровне рассмотрен ниже.

Аннотация морфологического уровня проводилась группой из семи аннотаторов, и была разделена на два этапа. На первом этапе каждый текст был предварительно аннотирован морфологическим анализатором. Затем два аннотатора, независимо друг от друга, проконтролировали правильность атрибутов lemma и tag. На втором этапе все несоответствия между этими двумя аннотаторами были разрешены третьим, контролирующим аннотатором. После окончания аннотирования поверхностно-синтаксического уровня была проведена еще одна ревизия, для проверки соответствия предлогов и падежей существительных, именного согласования и т.д.

2.2 Поверхностно-синтаксический уровень (ПСУ)

Здесь структура предложения представлена в виде ориентированного дерева с помеченными связями (ребрами) и узлами. Каждому элементу морфологического уровня соответствует узел поверхностно-синтаксического дерева, отношения между элементами выражены связывающими их ребрами. Тип отношения определяется типом ребра – большинство ребер отражают отношение зависимости, но есть и другие отношения, напр. координация, аппозиция, знаки препинания и др.

Каждому узлу приписывается шесть атрибутов. Атрибут id содержит однозначный в рамках PDT 2.0 идентификатор узла, который связывает его с глубинно-синтаксическим уровнем. Линейный порядок узлов отражается в атрибуте ord. Функция ребра по техническим причинам отображается в атрибуте afun у нижнего узла. Атрибуты is_member и is_parenthesis_root используются для указания на сочинительные конструкции и выражения в скобках. Атрибут m.rf связывает узел с соответствующим элементом на морфологическом уровне. Пример аннотации на ПСУ рассмотрен ниже.

Все данные ПСУ аннотировались группой из шести аннотаторов – сначала вручную, затем на основе предварительной автоматической аннотации. По окончании аннотирования проводились всевозможные контрольные тесты, найденные ошибки были вручную проверены и исправлены.

2.3 Глубинно-синтаксический (тектограмматический) уровень (ГСУ)

Структура ГСУ – дерево, где каждому узлу, кроме технического корня, присвоено 39 атрибутов. В зависимости от типа узла (атрибут nodetype) заполняется определенное подмножество этих атрибутов. Наибольший интерес представляют следующие атрибуты:

Атрибут functor – описывает тип ребра, ведущего от узла к его предку – зависимость или другое техническое отношение. Значениями этого атрибута могут быть функторы для актантов (ACT – агенс, PAT – пациенс, ADDR – адресат и др.), функторы корней независимых клауз (PRED – главный предикат предложения, DENOM – именной корень клаузы, PAR – корень выражения в скобках), функторы для корней сочинительных конструкций (CONJ – сочинительная конструкция, ADVS – противительная конструкция и др.), функторы места (LOC – где, DIR1 – откуда, DIR2 – каким путем, DIR3 – куда) и времени (TWHEN – когда, TTILL – до какого времени, TSIN – с какого времени, TPAR – в течение какого времени и др.) и другие. Всего на данный момент для аннотирования чешского языка используется 67 функторов, распределенных на 12 групп.

Атрибут t_lemma содержит имя лексемы на глубинно-синтаксическом уровне.

16 атрибутов используется для описания грамматических свойств узла. Эти атрибуты обозначены префиксом gram (напр., атрибут gram/sempos – семантическая часть речи, имеющий далее 19 значений: n.denot – семантическое существительное, adj.denot – семантическое прилагательное, v – глагол и т.д.; атрибут gram/verbmod содержит информацию о модальности предложения и т.п.)

Так как тектограмматическая структура, также как и ПСУ, основана на синтаксических зависимостях, для конвертирования поверхностно-синтаксических деревьев в предварительные глубинно-синтаксические были использованы автоматические методы. Все полученные таким образом деревья были затем вручную обработаны аннотаторами, которые дополнили большое количество недостающей информации и исправили ошибки.

062-Nedoluzhko:_Layout 1 18.05.2008 16:01 Page 403

2.3.1 Словарь моделей управления VALLEX

На ГСУ предикатам присваивается модель управления из связанного с TrEd-ом словаря валентностей **VALLEX**. Это электронный словарь, содержащий примерно 2730 лексем. Словарная статья включает как минимум одну модель управления с указанием обязательных актантов и их возможных синтаксических реализаций, а также с примерами их употребления. Например, представление глагола rozumět (понимать) в pdf-версии словаря выглядит так:

rozumět *impf* v

1 (*vyznat se; chápat*) ACT(1) PAT(3|zda|že|cont) ◊*rozumí úloze; nerozuměl, zda to má nebo nemá udělat; rozumíš už, co se stalo?; rozumí dobře anglicky; matka dceři rozumí* ✠ rcp: ACT-PAT; class: mental action

2 (*rozlišovat; znát*) ACT(1) ADDR(3) PAT(4|zda|že|cont) ¿MANN ◊*rozumí mu každé slovo; Rozumím vám správně, že rozpočet v parlamentě podpoříte? (ČNK)* ✠ rcp: ACT-ADDR; class: communication

3 idiom (*chápat*) ACT(1) PAT(4|že) EFF(7|pod+7) ◊*rozumí tím / pod tím přiměří* ✠ rfl: pass; class: mental action

При аннотировании ГСУ модель управления должна быть присвоена глаголам и отглагольным прилагательным на *-ní* (типа *koupaní – купание*) и *–tí* (*mýtí – мытье*). Предикативам – представителям других частей речи модель управления пока последовательно не присваивается.

Помимо грамматической структуры зависимостей, на тектограмматическом уровне имеется также информация об актуальном членении предложений и о кореференции, которая аннотировалась отдельно.

2.3.2 Актуальное членение

Аннотирация актуального членения основана на двух традиционных чешских концепциях: В. Матезиуса о теме-реме и контекстной связанности и Я. Фирбаса о функциональной перспективе предложения. В аннотации PDT 2.0 фиксируется контекстная связанность (данность, известность) узлов и функциональная перспектива предложения. Информация о тематических и рематических блоках должна автоматически высчитываться на основе этих данных. Контекстная связанность представлена значениями атрибута tfa (topic-focus articulation) – *t* (данное), *f* (новое) и *c* (контраст) и аннотируется вручную, отдельно для каждого узла. Атрибут deepord используется для обозначения глубинного порядка узлов, основанного на функциональной перспективе предложения. Таким образом, в глубинно-синтаксическом представлении порядок узлов слева направо обозначает степень их функциональной динамичности – от наименее к наиболее динамичному элементу.

2.3.3 Кореференция

В аннотации PDT 2.0 кореференция делится на грамматическую и текстовую. Другие случаи кореференции, такие как экзофорическая отсылка и отсылка к двум и более предложениям, рассматриваются отдельно. В качестве антецедента может выступать терминальный узел дерева, поддерево (отсылка на корень поддерева) или несколько поддеревьев (отсылка на корневые узлы этих поддеревьев)

В случае грамматической кореференции антецедент высчитывается на основании грамматических правил языка, к ней относится кореференция возвратных местоимений (в чешском языке возвратное местоимение – всегда отдельная клитика), относительных местоимений (напр. *человек, который* пьет; *в городе, где* мне так понравилось и др.) и др. Грамматическая кореференция практически никогда не переходит границ предложения, ее всегда можно представить как отсылку одного узла к другому, следовательно ее аннотирование легко автоматизируется.

Текстовая кореференция аннотируется в PDT 2.0 только в том случае, если в качестве анафорического члена выступают личные и посессивные местоимения третьего лица, указательное местоимение этот

062-Nedoluzhko:_Layout 1 18.05.2008 16:01 Page 404

в субстантивной функции и актуальный эллипсис этих местоимений, восстанавливаемый на ГСУ. Текстовая кореференция может легко переходить границы предложения, и ее определение часто зависит от знания контекста, поэтому ее аннотирование проводилось вручную.

Для аннотирования кореференции используется id антецедента, к которому отсылает id узла анафоры. Атрибуты coref_text.rf, и coref_gram.rf содержат id кореферентных узлов соответствующих типов. Атрибут coref_special несет информацию об особых случаях кореференции.

В настоящее время разрабатывается проект расширенного аннотирования кореференции, где текстовая кореференция будет дополнена случаями, когда в качестве анафорического повтора выступают другие части речи (прежде всего существительные – повторение данной ИГ, синонимы, гиперонимы и т.д.), но при этом сохраняется тождество референтов. Кроме того, планируется включить в аннотацию случаи т.наз. bridging anaphora, где референты антецедента и анафорического «повтора» уже не тождественны, но семантически связаны. Над этой темой сейчас работает автор данного доклада.

3. Пример предложения, аннотированного на трех уровнях аннотации в PDT 2.0

Některé kontury problému se však po oživením Havlovým projevem zdají být jasnější. – Некоторые контуры проблемы однако после оживлением выступления Гавела кажутся понятнее

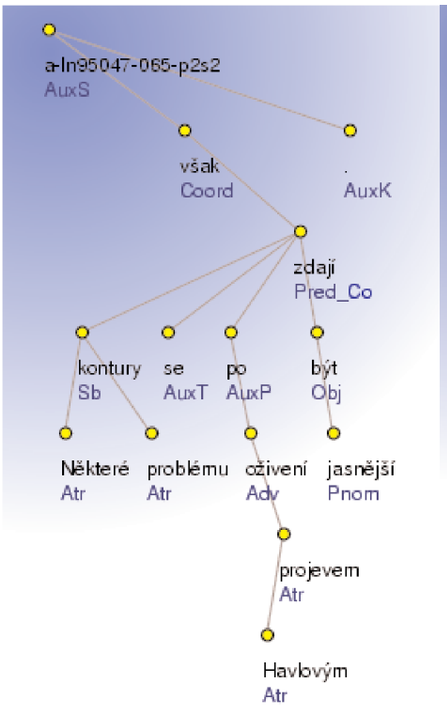
Některé	*kontury*	*problému*	*se*	*však*	*po*	*oživením*	*Havlovým*	*projevem*	*zdají*	*být*	*jasnější*
Некоторый adj. masc Npl	Контур noun, masc, Npl	Проблема masc, Gsg	возвр. «ся» pron.	Однако adv	По prep	Оживление noun,neutr, DSg	Гавлов adj-poss, masc, ISg	Выступления noun, neutr ISg	кажут(ся) verb, ind,act, praes. 3Sg	Быть inf	Ясный ср.степ.

3.1. Нулевой уровень слов

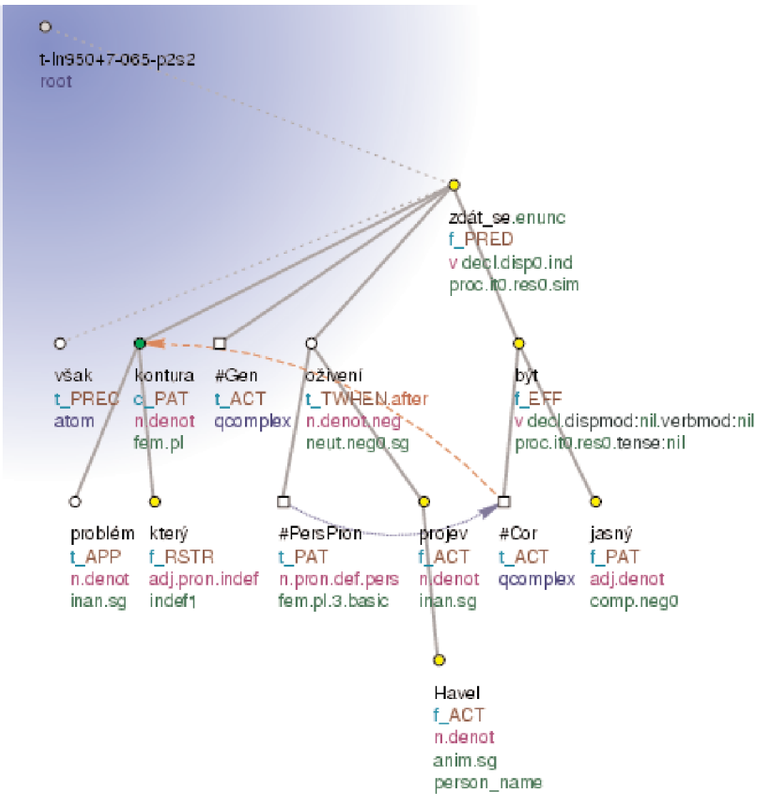
словоформа	**лемма**	**морфологический тег**
Některé	*některý*	PZFP1----------
kontury	*kontura*	NNFP1-----A----
problému	*problém*	NNIS2-----A----
se	*se_^(zvr._zájmeno/částice)*	P7-X4----------
však	*však*	J^----------
po	*po-1*	RR--6----------
oživení	*oživení_^(*3it)*	NNNS6-----A----
Havlovým	*Havlův_;S_^(*3el)*	AUIS7M---------
projevem	*projev*	NNIS7-----A----
zdají	*zdát*	VB-P---3P-AA---
být	*být*	Vf--------A----
jasnější	*jasný*	AAFP1----2A----
.	.	Z:----------

3.2. Морфологический уровень

062-Nedoluzhko:_Layout 1 18.05.2008 16:01 Page 405



3.3. Поверхностно-синтаксический уровень



3.4. Глубинно-синтаксический (тектограмматический) уровень

***

В недавнем прошлом проект PDT был дополнен Пражским арабским синтаксически аннотированным корпусом (Prague Arabic Dependency Treebank, http://www.ldc.upenn.edu) и параллельным чешско-английским корпусом (Prague Czech-English Dependency Treebank, http://ufal.mff.cuni.cz/pcedt). Арабский проект подтверждает, что разработанная на чешском языке система может работать и на типологически несходном языке. Синтаксически аннотированный параллельный чешско-английский корпус разрабатывается на основе аннотрирования вручную текстов из журнала Wall Street Journal, которые ранее использовались для корпуса Penn Treebank 3. В настоящее время параллельно аннотируется примерно 21600 предложений на английском языке и их переводы на чешский. Целью проекта является подготовка текстовой базы для обучения компьютера машинному переводу с чешского на английский и обратно.

Проект PDT имеет и более далеко идущие планы. Рассматриваются несколько вариантов: пополнение PDT разговорными текстами, детализация имеющейся аннотации (в основном в области аннотации кореференции, информационной структуры и дискурса), аннотация других типологически отличающихся языков, аннотация вручную глубинно-синтаксического уровня на параллельных чешских и английских текстах, разработка новых уровней аннотации и т.д. По большинству из этих проектов уже ведутся разработки.

Список литературы

1. Boguslavsky I., Grigorieva S., Grigoriev N., Kreidlin L., Frid N. Dependency Treebank for Russian: Concept, Tools, Types of Information. // In Proceedings of the 18th conference on Computational linguistics, Saarbrücken, Germany, 2000.

2. Čmejrek M., Cuřín J., Havelka J., Hajič J., Kuboň V. Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation, In 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal. Доступно на http://ufal.mff.cuni.cz/pcedt/doc/papers/lrec2004_pcedt.pdf. 2004.

3. Hajič J., Hajičová E., Hlaváčová J., Klimeš V., Mírovský J., Pajas P., Štěpánek J., Vidová-Hladká B., Žabokrtský Z. PDT 2.0 – Guide. UFAL & CKL, 2006 Доступно на http://ufal.mff.cuni.cz/pdt2.0/

062-Nedoluzhko:_Layout 1 18.05.2008 16:01 Page 406

4. Mikulova M. и кол. Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka Institute of formal and applied linguistics, Charles University, Prague, 2006.

5. Nědolužko A., Zpráva k anotování rozšířené textové koreference a bridging vztahů v Pražském závislostním korpusu. (Report about the annotation of the extended text-coreference and bridging relations in Prague Dependency Treebank.). Technical report. Institute of formal and applied linguistics, Charles University, Prague. 2007

6. Žabokrtský, Z.; Lopatková, M.: Valency Frames of Czech Verbs in VALLEX 1.0. // In Frontiers in Corpus Annotation. Proceedings of the Workshop of the HLT/NAACL Conference, pp. 70—77. 2004

063-Pavlova_A:_Layout 1 18.05.2008 16:01 Page 407

ЗНАЧЕНИЕ ПРОСОДИЧЕСКОЙ ИНФОРМАЦИИ В ЛЕКСИКОГРАФИЧЕСКОМ ТОЛКОВАНИИ ПОЛИСЕМИИ И ОМОНИМИИ

THE MEANING OF PROSODIC INFORMATION IN LEXIGORAFIC REPRESENTATION OF POLYSEMY AND HOMONIMY

Павлова А.В.*(anna.pavlova@gmx.de)*
SAP AG, Walldorf, Deutschland

Лексическая семантика слова в значительной мере предопределяет его сильную или слабую акцентную позицию, его тяготение к тематической или рематической части высказывания. Обнаружение связи между лексическим значением слова и его потенциальной ударностью или безударностью во фразе позволило бы уточнить лексикографическое описание значений. В частности, опора на такие связи способствовала бы более тонкой дифференциации значений полисемических лексем. К акцентному выделению тяготеет в первую очередь лексика с субъективно-оценочной, ретроспекивной и негативной семантикой. Однако имеется немало факторов, которые способны вступать в противоречие с просодическими характеристиками на уровне словаря: определенное коммуникативное задание (например, чистая повествовательность, причинность, событийность), фразеологическая связанность, однородные члены, употребление числительных. Если описывать потенциальное акцентное поведение того или иного слова в словаре, то необходимо по крайней мере в предисловии описать потенциальные же препятствия, которые в реальной речи способны свести на нет просодическую характеристику того или иного значения многозначного слова или одного из омонимов. Очевидно, что далеко не все слова в словаре (и даже, скорее всего, меньшинство) требуют подобной просодической информации. С другой стороны, иные значения столь тесно связаны с фразовой просодией, что пренебрегать этим фактом в лексикографии нельзя.

В литературе по актуальному членению предложения внимание в основном концентрируется на противопоставлении «данное – новое» в зависимости от контекста и на логических противопоставлениях: контрасте, антитезах, а также на роли логических частиц. Однако материал показывает, что во многих случаях решающим оказывается лексическая семантика определенных элементов предложения. Специальное внимание в этом докладе уделяется акцентному поведению слова в зависимости от употребления в том или ином из его значений. Этот аспект должен учитываться при составлении словарей.

Нас будет занимать исключительно восприятие письменной речи: ее декодирование интерпретатором на уровне актуального членения предложения. Теория актуального членения (ее называют также теорией коммуникативного или тема-рематического членения) насчитывает несколько десятилетий своей истории (первая статья на эту тему, работа В. Матезиуса, вышла в свет в 1947 году[1]. Эта теория сосредоточилась главным образом на объяснении коммуникативного членения порядком слов и контекстом, а также контрастом, логическими антитезами и ролью частиц в усилении или ослаблении рематической роли тех или иных элементов предложения)[2].

В работах по актуальному членению, наряду с подчеркиванием превалирующей роли порядка слов и контекста при определении места фразового ударения как метки ремы, отмечается большая степень свободы говорящего или озвучивающего письменную речь в расстановке фразовых ударений. Несомненно, произвольность трактовки коммуникативного членения, определения информационного центра и периферии высказывания воспринимающим письменную речь субъектом – это факт, который объясняет различия в озвучивании, например, одних и тех же художественных текстов разными чтецами. Тем не менее, представляется, что выделить и назвать определенные тенденции, позволяющие трактовать тема-рематический состав высказываний определенным способом с высокой степенью вероятности, возможно. В настоящем докладе

[1] См. [Mathesius 1947].
[2] См. [Ковтунова 1976], [Распопов 1970], [Селиверстова, Прозорова 1992], [Николаева 1982], [Николаева 1985].

063-Pavlova_A:_Layout 1 18.05.2008 16:01 Page 408

расстановка фразовых ударений во всех примерах основана на вероятностном подходе именно такого рода.

Непосредственных связей между лексическим значением наполняющих высказывание слов и их ролью в определении тема-рематического состава теория актуального членения до сих пор не установила, сосредоточившись, главным образом, на анализе соотношения между темой, ремой и контекстом (и – шире – экстралингвистическими знаниями о ситуации)[3]. Однако существует много случаев, которые, если и не опровергают представления о роли контекста и речевой ситуации как о с н о в н о г о фактора, определяющего коммуникативное членение предложения, то все же ставят ее под сомнение. Во-первых, в каждом письменном тексте имеется начальное предложение, для которого нет предшествующего контекста и которое, тем не менее, обретает коммуникативную структуру в сознании воспринимающего текст и прочитывается при необходимости вслух адекватным образом. И лишь изредка встречаются случаи, когда последующий контекст диктует необходимость исправить первоначально неверно понятое предложение. Во-вторых, имеется немало случаев, когда слово, являющееся элементом известной из контекста информации, оказывается в роли ремы. Так, в примере *(1) Ему принесли салат и сосиски. Он с жадностью накинулся на е´ду* – фразовое ударение во втором предложении приходится на информацию, известную из предыдущего предложения (сосиски и салат – это еда). Если следовать традиционным представлениям теории актуального членения, фразовое ударение, отмечающее рему, не должно приходиться на слово, обозначающее информацию, известную из непосредственно предшествующего контекста. Но при мысленном или внешнем озвучивании второго предложения из этого примера фразовое ударение с высокой степенью вероятности будет поставлено именно таким образом, как указано. Еще один пример: в комнате двое, Настя и ее начальник. Начальник что-то произносит. Далее следует фраза: *(2) Настя внимательно посмотрела ему в гла´за*. Вряд ли носителю русского языка, читающему эту фразу, захочется при ее понимании и произнесении поставить фразовое ударение на какое-либо другое слово, кроме последнего, хотя из непосредственного контекста ясно, что *глаза начальника* – известная информация, поскольку раз есть начальник, то у него есть и глаза[4]. Сопоставление этого примера с таким, как *(3) Настя увидела у´сталость в его глазах* – позволяет предположить, что определяющую роль в решении вопроса о расстановке фразовых ударений при понимании и письменной речи и произнесении фраз вслух играет не контекст, а лексическая семантика и/или связанный с ней синтаксический состав предложения. Поскольку начало и конец в обоих сопоставленных примерах лексически одинаковы, напрашивается вывод о том, что различие в акцентных структурах объясняется или расхождением каких-то семантических составляющих сказуемых *(внимательно) посмотреть* и *увидеть (усталость*, *злость, досаду, растерянность*), или падежной принадлежностью слова *глаза* в первом и во втором примерах.

(3а) Ему было ´холодно в одной рубашке. – (3б) Ему пришлось выйти в одной ру´башке. Представим себе, что эти предложения образуют собой абсолютное начало текста, то есть предшествующий контекст воспринимающему письменную речь неизвестен, знаний о речевой ситуации у него пока нет. Тем не менее, высока вероятность, что он произнесет данные фразы именно таким образом. Словосочетание *в одной рубашке* в первом предложении оформлено как старая информация, хотя она таковой формально быть не может, поскольку у нас нет контекста. Во втором примере то же словосочетание подается как новая информация. Что побуждает нас произвести актуальное членение этих предложений именно так, а не иначе? В отличие от первой пары примеров, где *глаза* имеют разную падежную принадлежность в первом и втором предложениях, словосочетание *в одной рубашке* в обоих предложениях второй пары примеров стоит в одном и том же – предложном – падеже. В роли кандидата на объяснение причины различного актуального членения остается семантическое различие между сказуемыми *было холодно* и *пришлось выйти* или/и различная синтаксическая роль словосочетания в предложном падеже в первом и втором предложении.

(4а) Мальчик сидел в углу и тихо ´вел себя – при таком порядке слов и фразовом ударении на *вел* фраза некорректна, для ее исправления требуется изменить порядок слов (*вел себя ’тихо*) или как минимум переместить фразовое ударение на *тихо*. *(4б) Мальчик сидел в углу и тихо ´пел* – корректная фраза. *(5а) Пособие позволяет впроголодь существо’вать* – фраза некорректна, в ней требуется изменить порядок слов и место фразового ударения: *(5б) Пособие позволяет существовать ´впроголодь.* Ср: *(5в) Пособие позволяет как-то (*вариант*: сносно) существо´вать* – это фраза правильная, некорректен был бы ее вариант с измененным порядком слов и смещением фразового ударения: *(5г) Пособие позволяет существовать ´как-то / ´сносно.* Предложения в сопоставляемых парах различаются исключительно лексическим значением одного-единственного элемента.

Эти примеры наглядно показывают, что а) фразовое ударение не безразлично к лексическому и/или

[3] «Нам неизвестны русские слова и конструкции, толкование которых имело бы непосредственное пересечение с определением р е м ы. Однако слова с рематической выделенностью существуют.» [Янко 2001:233]

[4] Можно было бы назвать эту информацию «данное», но этот термин уже занят: в теории актуального членения «данное» – синоним «темы», а «новое» – синоним «ремы».

063-Pavlova_A:_Layout 1 18.05.2008 16:01 Page 409

синтаксическому составу предложения и б) контекст способен уступать свою главную роль определителя места фразового ударения лексическому и/или синтаксическому составу предложения.

К сожалению, объем доклада не позволяет формализовать и описать правила или, точнее, тенденции, которыми руководствуется интерпретатор письменного высказывания, прочитывая его так или иначе с точки зрения его акцентной структуры в зависимости от его лексического состава. Эти тенденции поддаются известной алгоритмизации, но это уже материал для книги, так как тенденций этих много. Некоторые из них вступают в отношения конкуренции, в которых одна из них одерживает верх. Их описание требует тщательного подбора пар примеров различного коммуникативного членения, которые отличались бы друг от друга только одной составляющей, иначе достоверность тенденции оказалась бы под сомнением. В рамках данной статьи приходится ограничиться наблюдениями лишь над некоторыми фактами актуального членения и вытекающими из этих наблюдений рекомендациями[5].

Следующие примеры подтверждают справедливость предположения о роли прежде всего лексического состава предложения в принятии решения о его коммуникативном членении: *(6а) Тоска о´ставила меня. – (6б) То´ска накатила на меня; (7а) Он бы ´рад разрыву Наташи с мужем. – (7б) Приходите, буду рад вам по´мочь; (8а) Моя жена – превосходная кули´нарка. – (8б) Моя жена – по´средственная кулинарка; (9а) У нее были свои ´слуги. – (9б) У нее были сво´и взгляды; (10а) У него были знаменитые ро´дители. – (10б) У него б´ыли свои заслуги перед наукой (но человек он был малосимпатичный); (11а) У нас есть полный о´бед. – (11б) У нас ´есть о чем поговорить.*

Некоторые пары примеров обнаруживают различия в лексической семантике слова в зависимости от его ударной или безударной позиции. Так, *свои* в (9а) означает ‚собственные', а в (9б) – ‚своеобразные'. *Рад* в (7а) означает ‚радовался', а в (7б) выражает формулу вежливости: это не указание на эмоцию, а метка приветливого тона, призванного помочь преодолеть возможные сомнения или застенчивость того, кому предлагается помощь. *Были* в (10а) выражает объективное отношение принадлежности, а в (10б) – уступительность, субъективную оценку объекта и намек на продолжение, содержащее некоторое противоположное утверждение об объекте *(‚ну да, я признаю, что он в известном смысле неплохой ученый, но человек он при этом малосимпатичный').* Безударное *есть* в (11а) выражает факт наличия, а ударное *есть* в (11б) в сочетании с *о чем поговорить* содержит субъективную оценку: уверенность говорящего в том, что на разговор стоит потратить время.

Справедлива и обратная зависимость: то или иное лексическое значение определяет, ставить ли на слове фразовое ударение или нет: *(12) Я запустил процесс приготовления еды.* – В зависимости от того, означает ли *запустил ‚начал'* или *‚перестал'*, глагол во фразе соответственно безударен или несет на себе фразовое ударение; *(13) Я просто смотрю на вещи.* – В зависимости от того, означает ли *просто* оценку *(‚отношусь к миру стараясь его излишне не усложнять')* или характеризует действие *смотрю (‚смотрю и больше ничего не делаю'), просто* в предложении ударно или безударно. Попутно меняется лексическое значения слова *вещи*: в первом имеется в виду отображение мира в чьем-то сознании, во втором – конкретные вещи, например, в комнате. *(14) В этой семье умеют хорошо печь.* – Если *умеют* означает способность, умение – то ударение на *печь*, а если это похвала *(‚понимают толк в выпечке'),* то ударение на *умеют*. *(15) Посмотрите на меня.* – Если глагол *посмотрите* означает *‚взгляните'*, то он под ударением, а если *‚берите с меня пример',* то он безударен.

Наблюдение над зависимостью некоторых лексических значений от ударности или безударности подводит нас к заключению о том, что перевод также не может быть безразличен к ацентной позиции лексемы во фразе. Так, перевод на немецкий язык слова *свои* в паре примеров (9а) и (9б) явно разный: *(16а) Sie hatte ihre eigenen 'Dienstboten. – (16б) Sie hatte 'eigenartige (ausgefallene, merkwürdige, ungewöhnliche) Ansichten.* Пару предложений с *есть* в (11а) и (11б) переведем как: *(17а) Wir haben (oder bieten) heute ein komplettes Me´nu. – (17б) Wir haben ge´nug zu besprechen*.

Толковые словари по-разному представляют многозначность. Так, для *свой* словарь Ожегова-Шведовой описывает оба значения, соответствующие (9а) и (9б). А вот на разницу в значениях *есть, быть*, манифестируемую позицией фразового ударения в (10а) и (10б) и проявляющуюся при переводе, словарь Ожегова-Шведовой не указывает. Необходимость различного перевода на другие языки обнаруживаем и в таких примерах, как: *(18а) Sie ist einfach in ihrem Ele´ment. – (18б) Sie ist ´einfach in ihren Lebensansichten.* Первое, безударное, *einfach* в (18а) означает ‚просто, попросту' и с точки зрения грамматической принадлежности является наречием или частицей, а в рамках синтаксической структуры предложения это модальное слово, выражающее отношение говорящего к содержанию какого-то предыдущего высказывания собеседника. Ударное *einfach в (18б)* означает ‚наивный', ‚неглубокий', ‚примитивный'; грамматически это прилагательное,

[5] Интересные наблюдения о «словах в темы и словах ремы» содержатся в 3-ей главе книги [Янко 2001]. См. тж. описание рематических и тематических глаголов в [Павлова 1987], описание семантики прилагательных в связи с их ролью в актуальном членении предложения в [Скорикова 1982]. Попытка обобщения семантических факторов, влияющих на рематичность или тематичность тех или иных слов, групп слов и словосочетаний содержится в [Павлова 2007].

063-Pavlova_A:_Layout 1 18.05.2008 16:01 Page 410

синтаксически – предикатив. Словарь Wahrig содержит описание обоих значений, но, разумеется, без ссылки на разное акцентное поведение слова в одном и другом его значении в высказывании.

(19а) Die meisten der bedeutenden Propheten, angefangen von Amos, befassen sich ´wenig mit theologischen Spekulationen (E. Fromm). – (19б) Die meisten der bedeutenden Propheten, angefangen von Amos, befassen sich ein wenig mit theologischen Spekula´tionen. (19а) переводится так: *Большинство известных пророков, начиная с Амоса, ´мало / не´много размышляли о теологии,* (19б) – как *Большинство известных пророков, начиная с Амоса, немного размы´шляли о теологии*. Слово *немного* в первом предложении означает ‚мало', во втором – 'слегка', ‚в некоторой степени'. В словаре Ожегова-Шведовой оба значения *немного* приводятся через запятую, как одно. Их стоило бы разделить: не только акцентная и тема-рематическая структура предложения зависит от значения *немного*, но и перевод будет осуществлен по-разному (*wenig* или *ein wenig*) в зависимости от того, означает *немного* ‚мало' или ‚слегка'.

Изредка попадаются примеры «акцентной омонимии» как метки многозначности лексемы в условиях полного совпадения прочих элементов предложения. Например: *(20) После отрицательного ответа папы Римского на свое прошение он оставил свою конфессию.* В зависимости от того, как будет истолковано значение *оставил* – как ‚покинул', ‚отказался' или как ‚сохранил', – фразовое ударение придется на *оставил* в первом случае или на *свою* во втором. Полное прояснение положения дел дает только последующий контекст, из которого выясняется, что «он» перешел из католичества в лютеранство. В зависимости от значения слова *оставил* в его ударной или безударной позиции, естественно, фразы переводятся по-разному: *er gab seine Konfession auf* / *er blieb bei seiner Konfession*. Словарь Ожегова-Шведовой содержит описание обоих значений глагола *оставлять*.

(21а) Он поспешил уй´ти. – (21б) Он поспе´шил уйти. В (21а) *поспешил уйти* означает ‚поспешно, быстро ушел', в (21б) содержится модальная оценка говорящего (осуждение, неодобрение): ‚зря он ушел'. От поспешности как таковой в значении сказуемого второй фразы ничего не осталось. Словарь Ожегова-Шведовой включает описание только одного значения глагола *поспешить*.

(22а) Я специально не ´буду заходить. – (22б) Я специ´ально не буду заходить. Перевод этих предложений выглядит так: *Ich werde nicht extra vorbeikommen (‚отдельно, только для чего-л.'). – Ich werde absichtlich nicht vorbeikommen (‚намеренно').* Словарь Ожегова-Шведовой не приводит разных значений слова *специально*.

(23а) Du be´greifst langsam. – (23б) Du begreifst ´langsam. Фразовое ударение приходится на *langsam* или минует его в зависимости от того, означает ли *langsam* ‚постепенно', как в (23а), или ‚небыстро, медленно', как в (23б). Разумеется, в принятии решения о расстановке фразовых ударений в таких случаях примерах не обойтись без контекста, проявляющего значение *langsam*. Словарь Wahrig описывает оба значения. Также для пары примеров *(24а)* '*Geh ruhig. – (24б) Geh ´ruhig*, в которой первая фраза означает ‚*ступай себе'*, а вторая ‚*иди спокойно, не спеши'* – словарь Wahrig содержит описание обоих значений.

Любопытно, что русское *спокойный* обнаруживает аналогичное расхождение в значениях: *(25) Он спокойно ел.* Здесь *спокойно* может означать ‚он ел себе', то есть ел невозмутимо, не проявляя никаких эмоций, а может – ‚неспешно, неторопливо'. Только во втором своем значении это слово при определенных контекстуальных условиях может оказаться под фразовым ударением и выступить в роли ремы: *(25а) Он спо'койно ел.* Скорее всего, для подкрепления его роли ремы и облегчения задачи правильной расстановки фразового ударения в русском языке на письме автор изменит порядок слов: *(25б) Он ел спо´койно.* Значение, выражаемое словом *себе*, делает это слово принципиально безударным, при этом значении его грамматическая функция приближена к функции частицы. Между тем, словарь Ожегова-Шведовой не упоминает для *спокойный* значения ‚невозмутимый', ‚без выражения эмоций', в отличие от словаря Wahrig в отношении его немецкого эквивалента *ruhig*.

Лексическая семантика слова в значительной мере предопределяет его сильную или слабую акцентную позицию, его тяготение к тематической или рематической части высказывания и, следовательно, влияет также на положение слова во фразе. Это демонстрируют многие приведенные выше примеры. Словарное значение лексемы тесно связано с ее потенциальной способностью выступать в качестве элемента темы или акцентного центра высказывания – ремы. Не распространяя это утверждение на весь лексикон, отметим однако, что обнаружение связи между лексическим значением слова и его потенциальной ударностью или безударностью во фразе позволило бы уточнить лексикографическое описание значений. В частности, опора на такие связи способствовала бы более тонкой дифференциации значений полисемических лексем.

Таким образом, взаимодействие фразового ударения с лексической семантикой конкретных многозначных слов (или омонимов, развившихся из полисемии) имеет непосредственное отношение к лексикографии. Множество примеров свидетельствует в пользу положения о необходимости снабжать лексикографические толкования информацией о просодическом «поведении» слова во фразе[6].

[6] Мысль о необходимости этого высказывается в [Апресян 1995, 178-197] и [Апресян 1990] , а также в [Скорикова 2002].

063-Pavlova_A:_Layout 1 18.05.2008 16:01 Page 411

Очевидно, что фразовое ударение при произнесении высказываний в устной речи обладает способностью к передвижению сем в рамках одного значения или к замене одного значения многозначного слова другим. При восприятии письменной речи, наоборот, место фразового ударения часто диктуется и определяется тем или иным выбором значения многозначного слова. Однако далеко не все значения многозначных слов столь тесно связаны с фразовым ударением. Можно привести множество примеров, когда фразовое ударение остается безразличным к смене значения многозначного слова. Так, многозначность глагола *снять* в словосочетаниях *снять пальто, снять начальника, снять квартиру, снять крышку, снять сцену на пленку, снять кандидатуру* и *снять диагноз* никак не отражается на отношении этого глагола к фразовому ударению, за исключением двух последних значений, ср. *(26а) Гость снял паль′то. – (26б) В нашем отделе сняли на′чальника. – (26в) Спустя месяц они сняли ква′ртиру. – (26г) Я сняла со сковородки ′крышку. – (26д) Немцов ′снял свою кандидатуру. – (26е) Через год врачи ′сняли диагноз.* Таким образом, смена значения далеко не всегда влечет за собой смену места фразового ударения. К акцентному выделению тяготеет в первую очередь лексика с субъективно-оценочной, ретроспекивной и негативной семантикой. Помимо семантических ограничений, существует немало дополнительных факторов, которые способны вступать в противоречие с любыми просодическими характеристиками на уровне словаря. Так, акцентуации слова в силу его лексического значения способны противостоять такие факторы, как определенное коммуникативное задание (например, чистая повествовательность, причинность, событийность, употребление повелительного наклонения), фразеологическая связанность, однородные члены, употребление числительных и некоторые другие. Если описывать потенциальное акцентное поведение того или иного слова в словаре (а автор настоящей работы приверженец такой практики), то необходимо по крайней мере в предисловии описать потенциальные же препятствия, которые в реальной речи способны свести на нет просодическую характеристику того или иного значения многозначного слова или одного из омонимов. Очевидно, что далеко не все слова в словаре (и даже, скорее всего, меньшинство) требуют подобной просодической информации. С другой стороны, иные значения столь тесно связаны с фразовой просодией, что пренебрегать этим фактом в лексикографии нельзя. Это означает давать неполную семантическую информацию о лексическом составе языка. Кроме того, акцентуация или, наоборот, безакцентная позиция того или иного слова в парах идентичных или минимально различающихся предложений – это потенциальный дополнительный источник информации о словарной полисемии.

Список литературы

1. Mathesius V. O tak zvaném aktuálnim členěni větném//Čeština a obecný jazykozpyt. Praha, 1947.

2. Wahrig. Deutsches Wörterbuch. 7. Auflage. München, 200

3. Апресян Ю. Д. Типы лексикографической информации об означающем лексемы // Типология и грамматика. М., 1990.

4. Апресян Ю. Д. Изб. тр.: Интегр. описание яз. и систем. лексикография. М., 1995. С. 178–197.

5. Матезиус В. О так называемом актуальном членении предложения//Пражский лингвистический кружок. М., 1967.

6. Ковтунова И. И. Порядок слов и актуальное членение предложения. М., 1976.

7. Николаева Т. М. Семантика акцентного выделения. М., 1982.

8. Николаева Т. М. Функции частиц в высказывании. М., 1985.

9. Ожегов С., Шведова Н. Толковый словарь русского языка. М., 2006.

10. Павлова А. В. Акцентная структура высказывания в ее связях с лексической семантикой: Автореф. канд. дис. Л., 1987.

11. Павлова А. В. Интерпретация акцентной структуры высказывания при восприятии письменной речи // Acta Linguistica Petropolitana. Труды Ин-та лингвистических исследований РАН. Т. III. СПб, 2007 (в печати).

12. Распопов И. П. Строение простого предложения в современном русском языке. М., 1970.

13. Селиверстова О. Н., Прозорова Л. А. Коммуникативная перспектива высказывания//Теория функциональной грамматики. Субъектность. Объектность. Коммуникативная перспектива высказывания. Определенность/неопределенность. СПб., 1992.

14. Скорикова Т. П. Функциональные фозможности интонационного оформления словосочетания в потоке речи. М., 1982.

15. Скорикова Т .П. Принципы описания акцентогенных свойств лексем // Проблемы фонетики IV. М., 2002. С. 181–190.

16. Янко Т. Е. Коммуникативные стратегии русской речи. М., 2001.

064-Paducheva:_Layout 1 18.05.2008 16:02 Page 412

РЕЖИМ ИНТЕРПРЕТАЦИИ КАК КОНТЕКСТ, СНИМАЮЩИЙ НЕОДНОЗНАЧНОСТЬ

REGISTER OF INTERPRETATION AS DISAMBIGUATING CONTEXT

Падучева Е.В. *(elena708@gmail.com), ВИНИТИ РАН*

В современной семантике фокус внимания смещается от описания отдельных значений языковых единиц к описанию семантических переходов и контекстов, обуславливающих сдвиг значения. Одним из таких контекстов является режим интерпретации. В докладе приводятся примеры эгоцентрических грамматических категорий, слов, конструкций, которые имеют в разных режимах разную интерпретацию.

По мере того, как совершенствуются наши методы описания значения, всё яснее дает о себе знать многозначность: многозначность в природе языка. Особенно это касается регулярной многозначности (Апресян 1974). Поэтому в современной семантике фокус внимания всё явственнее смещается от описания отдельных значений языковых единиц к описанию семантических переходов и контекстов, обуславливающих сдвиг значения.

Одним из таких контекстов стал в последнее время РЕЖИМ ИНТЕРПРЕТАЦИИ; это контекст, необходимый для понимания значения ЭГОЦЕНТРИЧЕСКИХ (в частности, ДЕЙКТИЧЕСКИХ) элементов языка – грамматических категорий, слов, конструкций. Семантика эгоцентрического элемента (ЭЭ) предполагает присутствие в ситуации некоего субъекта – говорящего или его аналога.

Можно различить три режима интерпретации. Основным является РЕЧЕВОЙ, или ДИАЛОГИЧЕСКИЙ режим. Ему соответствует КАНОНИЧЕСКАЯ КОММУНИКАТИВНАЯ СИТУАЦИЯ (Lyons 1979: 579ff), когда есть говорящий и слушающий, которые связаны единством места и времени; имеют общее поле зрения; могут видеть друг друга и жесты друг друга, и т.д. Речевому режиму противопоставлены НАРРАТИВНЫЙ и ГИПОТАКСИЧЕСКИЙ.

Принято различать первичный и вторичный дейксис (ср. Апресян 1986). Под вторичным дейксисом понимаются такие явления, как дейксис художественного повествования; дейксис пересказа (так сказать, «третичный»); дейктическая проекция по Лайонзу (т.е. дейксис в неканонической, но все-таки диалогической ситуации); сюда же следует, видимо, отнести также дейксис в гипотаксическом контексте. Я применяю противопоставление «первичный–вторичный» не только к контекстам употребления дейктических и, вообще, эгоцентрических элементов, но и к самим этим элементам. ПЕРВИЧНЫЕ ЭГОЦЕНТРИКИ – это слова и категории, которые полностью реализуют свой смысл только в условиях канонической речевой ситуации и ориентируются только на полноценного говорящего (а в не-канонической ситуации либо не употребляются, либо меняют свое значение); а ВТОРИЧНЫЕ могут, не меняя значения, ориентироваться не на говорящего, а на другое лицо.

Моя задача – показать, что режим интерпретации является контекстом, который необходим для раскрытия значения эгоцентрических элементов и снятия их неоднозначности. Чтобы дать общее представление о режимах интерпретации, мне придется частично повторить примеры, приводившиеся в Падучева 1986, 1996, 2004.

1. Первичные и вторичные эгоцентрики

Говорящий может присутствовать в семантике языковых единиц в разных ипостасях. Прежде всего – как субъект РЕЧИ. А кроме того, как субъект ДЕЙКСИСА, субъект ВОСПРИЯТИЯ и субъект СОЗНАНИЯ (т.е. модальности, мнения, оценки, эмоции, и тому подобное).

Примером первичного эгоцентрика могут служить местоимения *я, ты, вон <там>*. Примером вторичного – неопределенные местоимения; так, *какой-то* предполагает субъекта неопределенности-незнания – им может быть как говорящий, так и другое лицо. Глаголы *послышаться, показаться*, *раздаться, доноситься* предполагают наличие в ситуации субъекта восприятия, иначе – Наблюдателя. Роль Наблюдателя тоже может выполнять как говорящий, так и другое лицо. Предложение (1а) – это знаменитый пример Наблюдателя из Апресян 1986, лежащий в основе понятия режима интерпретации:

(1) а. На дороге *показался* всадник [субъект восприятия, предположительно, говорящий];

б. Иван шел к морю. Вдруг на дороге *показался* всадник [предположительно, субъект восприятия – Иван].

Глаголы *кривляться, ломаться* предполагают субъекта оценки; *сейчас, неподалеку* – субъекта дейксиса; прилагательные и наречия *ясно, грустно, досадно, надо* – субъекта сознания, ср. примеры, которые приводятся

064-Paducheva:_Layout 1 18.05.2008 16:02 Page 413

в связи с подразумеваемым субъектом сознания в статье В.В.Виноградова о «Пиковой даме» (Виноградов 1976):

(2) а. Долгая зимняя ночь прошла *незаметно*;

б. В одном увидел он черноволосую головку, наклоненную, *вероятно*, над книгой или над работой;

в. Неведомая сила, *казалось*, привлекала его к нему <дому графини>.

Все эти эгоцентрики вторичные.

При описании семантики языковых единиц лингвист обычно имеет в виду речевой контекст. Между тем, если контекст нарративный, то значение ЭЭ может меняться. Особенно это касается первичных эгоцентриков: вторичный эгоцентрик, по определению, имеет в нарративном контексте то же значение, что в речевом. Итак.

Вторичный эгоцентрик в нарративном контексте не меняет значения, но допускает две ОРИЕНТАЦИИ – персональную, как в (3б), и повествовательную, как в (3а):

(3) а. Нет! Мастер ошибался, когда говорил Иванушке в больнице в тот час, когда ночь перевалилась через полночь, что она позабыла его. Она его, *конечно*, не забыла; [субъект уверенности – повествователь, который знает, как было дело; **повествовательная** интерпретация-ориентация]

б. Она сделала все, чтобы разузнать что-нибудь о нем, и, *конечно,* не разузнала ровно ничего. [пессимизм может принадлежать и Маргарите, **персональная** интерпретация]

Чаще, все-таки, субъектом уверенности является персонаж. И субъектом незнания тоже чаще бывает персонаж, чем повествователь. (Ср. незнающего повествователя у Гоголя и ненадежного повествователя в романе Набокова «Пнин».) Неоднозначность ориентации может, впрочем, возникать и в речевом контексте.

Первичный эгоцентрик, который требует канонической коммуникативной ситуации и полноценного говорящего, в контексте нарратива **либо**

а) не употребляется вообще (*... испекли мы каравай вот такой вышины*), **либо**

б) имеет другой смысл (ср. *Вот твоя книга*! = ‘здесь’ и *Вот пошел Иван в лес*), **либо**

в) выводит за пределы диегетического пространства текста, выявляя внеположного этому пространству повествователя.

Есть еще одна возможность для первичного эгоцентрика в нарративе – несобственная прямая речь (НПР); это когда первичный эгоцентрик оказывается в распоряжении персонажа 3 лица; получаются, как говорит А.Банфилд, unspeakable sentences:

(4) Князю *бы не хотелось* расставаться с этим крестом. [ср. *Мне бы не хотелось*] (Достоевский. Идиот)

(5) *Лично ей* это было безразлично. (Р.Киреев. Ровно в семь у метро) [ср. *Лично мне* это безразлично]

Итак, чтобы говорить о значении эгоцентриков, надо идентифицировать режим интерпретации. Ниже будет показано на примерах, как режим интерпретации снимает неоднозначность ЭЭ. Предварительно, несколько уточнений.

1. Говоря о нарративе, мы до сих пор имели в виду повествование от 3 лица; прош. в примере (2) – время повествователя. Употребление эгоцентриков в повествовании от 1 лица отличается от третьеличного нарратива незначительно. В нарративе 1 лица употребляются, не выводя за пределы текстового пространства, эгоцентрики (в том числе – первичные), ориентированные на субъект РЕЧИ. Примеры из Национального корпуса русского языка (сокращенно –НКРЯ, адрес в Интернете – www.ruscorpora.ru). Оборот *по правде говоря* вполне уместен в (6), где есть *я*; а пример (7) этот оборот заставляет идентифицировать как несобственную прямую речь; так что (6) возможно в обычном речевом контексте, а (7) – специфический нарратив:

(6) <…> доставать [ящик с игрушками] мне и самой, *по правде говоря*, не хотелось.

(7) Тане, *пожалуй*, было интереснее общаться с Виталиком, поскольку он склонялся к медицине и у них было больше общих тем, но, *по правде говоря*, в качестве кавалеров её гораздо больше устраивали посторонние мальчики [Л. Улицкая. Путешествие в седьмую сторону света].

Что же касается первичных эгоцентриков, апеллирующих к времени и месту говорящего, дейктических, то они в повествовании от 1 лица невозможны совершенно так же, как в третьей форме, поскольку у рассказчика и читателя нет общего пространства, поля зрения и момента речи. Пример из Гуковский 1959: 115 [*здесь* = ‘в том месте, где находится повествователь’]:

(8) Впрочем, я думаю, что не имеет ли самый воздух в Малороссии какого-то особенного свойства, потому что если бы *здесь* вздумал кто-нибудь таким образом накушаться, то, без сомнения, вместо постели очутился бы лежащим на столе. (Гоголь. «Старосветские помещики»)

2. Понятие режима интерпретации, т.е. разграничение речевого и нарративного режима, возникло на базе введенного Бенвенистом разграничения «плана повествования» и «плана речи». Однако режим интерпретации – понятие более изощренное. Интерпретация ЭЭ не только идентифицирует режим, но и выявляет ориентир. В одном и том же тексте **как правило** возможны разные режимы интерпретации для разных эгоцентриков; и уж заведомо разные ориентиры. Т.е. текст не принадлежит весь к какому-то одному плану. Так что надо уметь интерпретировать отдельные ЭЭ текста. Идентификация текста в целом как относящегося к тому или иному типу может быть разве что статистическая, ср. терминологические неудобства, которые возникают из-за этого в Шмид 2000.

064-Paducheva:_Layout 1 18.05.2008 16:02 Page 414

В.В.Виноградов заключает в своей статье о «Пиковой даме» раздел под названием “Образ автора в композиции «Пиковой дамы»” словами: «Итак, мало того, что структура субъекта подвижна, что она двойственна по отношению к сферам сознания двух персонажей, но она вообще лишена точно очерченной субъективности. В самом субъекте как форме повествования заложена возможность приятия автором и мира Германна, и мира Лизы». Иначе это наблюдение В.В.Виноградова можно выразить так: повествователь, Лиза, Германн и даже старая графиня приблизительно равны в своем праве быть подразумеваемыми субъектами эгоцентриков этого текста.

3. Обращение к гипотаксическому режиму интерпретации ЭЭ позволяет ясно понять пример из Успенский 1970/2000: 79:

(9) … ей так хотелось поскорее, полегче, перелить из себя свое знание в *ребенка*, уже боявшегося, что вот-вот *тетя* рассердится, что она при малейшем невнимании со стороны *мальчика* вздрагивала, торопилась, горячилась, … (Война и мир).

В этом отрывке основной субъект сознания – княжна Марья; Б.А.Успенский отмечает номинацию *тетя* восклицательным знаком. В самом деле, тут в пределах одной фразы – двойной переход от одного ориентира к другому, хотя обычно субъект сознания на протяжении какого-то отрезка текста один и тот же. А дело в том, что номинация *тетя* возникает в гипотаксическом контексте: субъект номинации – ребенок «боявшийся». Только поэтому двойная смена субъекта сознания не означает отклонения от нарративных норм, т.е. нарративной неудачи.

В следующем фрагменте из «Анны Карениной» двойная смена ориентира – носителя точки зрения (в квадратных скобках указан предполагаемый для данного фрагмента текста Субъект сознания):

(10) [повествователь: Она *тоже* не спала всю ночь и все утро ждала его.] [Кити: Мать и отец были *бесспорно* согласны и счастливы ее счастьем. Она ждала его. Она первая хотела объявить ему свое и его счастье. Она готовилась одна встретить его и радовалась этой мысли, и робела, и стыдилась, и сама не знала, что она сделает. Она слышала его шаги и голос и ждала за дверью, пока уйдет mademoiselle Linon. Mademoiselle Linon ушла.] [повествователь: Она, не думая, не спрашивая себя, как и что, подошла к нему и сделала то, что она сделала.]

Слово *тоже* – эгоцентрик, поскольку пресуппозиция ‘Левин не спал всю ночь’ должна кому-то принадлежать. Она может принадлежать только повествователю, поскольку Кити не может знать про то, как провел ночь Левин. Дальше идет повторное описание сцены с mademoiselle Linon. Этот повтор мотивирован, очевидно, тем, что меняется персонаж – носитель точки зрения: раньше это был Левин, теперь – Кити. Но последняя фраза – снова от лица повествователя. Неясно, что это: прием или нарративная неудача.

Ниже следуют примеры употребления эгоцентриков в речевом и нарративном режиме. Рассматриваются три вида ЭГОЦЕНТРИКОВ: эгоцентрические грамматические категории, слова и конструкции. Показано, как режим интерпретации выявляет и снимает неоднозначность и какую роль в семантике вторичного эгоцентрика играет ориентир.

2. Грамматические категории

Пример 1. Главным примером первичного эгоцентрика в русском языке является грамматическое время (именно на анализе системы французских времен построена теория планов речи у Бенвениста (1959/1974)). В самом деле, между значением форм времени в речевом режиме и в нарративе вопиющая разница. Согласно общепринятым определениям, настоящее время означает одновременность МОМЕНТУ РЕЧИ; прошедшее выражает предшествование, будущее относит событие к некоему последующему моменту. Между тем, форма прош. времени несов. вида, которая в речевом режиме обозначает **предшествование** (моменту речи), в нарративном выражает **синхронность** (текущему моменту текста, т.е. текстовому времени):

(1) а. Кто *сидел* на моем маленьком стульчике и сломал его? (Толстой. Три медведя) [прямая речь; т.е. речевой режим; форма прош. выражает **предшествование**]

б. Старая графиня*** *сидела* в своей уборной перед зеркалом. Три девушки *окружали* ее. (Пушкин. Пиковая дама) [нарративный режим; прош.время выражает **синхронность**]

Прош. время во фразе (2) (почти цитата из романа А.Фадеева «Разгром») имеет в речевом и нарративном режиме понимания, существенно различные между собой, в речевом это долженствование в настоящий момент, в нарративном – сожаление о невыполненном долге:

(2) Надо было исполнять свои обязанности.

Пример 2. Повествовательная и персональная интерпретация форм времени. Форма наст. в примере (1) – это несобственная прямая речь.

(1) Беликов нервно засуетился и стал одеваться быстро, с выражением ужаса на лице. Ведь это в первый раз

064-Paducheva:_Layout 1 18.05.2008 16:02 Page 415

в жизни он *слышит* такие грубости. (Чехов. Человек в футляре) [настоящее персонажа; НЕ наст. историческое]

(2) Где же *были* ее краски? <…> Да, краски. Она оставила их в гостиной *вчера вечером*. (W.Woolf. To the Lighthouse)[прош. время – время повествователя]

Пример 3. В речевом режиме русская форма совершенного вида и англ. Present Perfect могут быть равны по смыслу:

(1) а. My uncle *has broken* his leg [событие в прошлом; и состояние, актуальное в момент речи];

б. Мой дядя *сломал* ногу [событие в прошлом; и состояние, актуальное в момент речи].

Однако Present Perfect первичный эгоцентрик (он может быть ориентирован **только** на момент речи), а СВ – вторичный, и сочетается с обстоятельством времени. В контексте обстоятельства перфектность СВ, т.е. идея сохранения результата в момент речи, пропадает – пропадает и сходство с англ. перфектом:

(2) а. Ко мне *приехали* родственники [событие; и состояние, актуальное в момент речи];

б. Ко мне сейчас *приехали* родственники [состояние, актуальное в момент речи].

в. Ко мне вчера *приехали* родственники [событие]; ср. *My relatives *have come* yesterday.

В этой связи вызывает сомнение анализ, предложенный в Апресян 1986 для примеров (3а) и (3б) – «в случае сов. вида говорящий мыслит время события как образующее единое целое со своим временем, т.е. тем временем, в котором он мыслит себя»; отсюда выводится свойственный сов. виду эффект сохранения результата в настоящем:

(3) а. Я *прочел* «Войну и мир» в раннем детстве;

б. Я *читал* «Войну и мир» в раннем детстве;

в. Я *прочел* «Войну и мир».

На самом деле, сов. вид имеет перфектную семантику только в контексте предложения (3в), без обстоятельства, а различие между (3а) и (3б) можно усмотреть разве что в эксплицитности указания на доведения действия до конца.

Ю.С.Маслову принадлежит открытие того замечательного факта, что форма причастия на *н*/*т* без связки – это, в современном языке, не наст. время пассива, а особый вид: **статальный перфект** (Маслов 1983, 1987). В речевом режиме статальный перфект обозначает не событие и состояние, а только состояние. Причем в речевом режиме это состояние, которое актуально в **момент речи. Е**ще в XIX веке статальный перфект не сформировался, отсюда неправильное понимание школьниками строчек из Лермонтова:

(4) – Скажи-ка, дядя, ведь не даром / Москва, спаленная пожаром, /Французу *отдана*? (пример из Падучева 2004: 498)

В полном соответствии с этим анализом получаем аномалию в примере (5), где статальный перфект попадает в контекст ЦЕПОЧКИ. В самом деле, здесь глагол в позиции первого звена в цепочке означает, что состояние актуально в момент речи, а второй глагол противоречит такому пониманию:

(5) *Телега *сломана* и после починена (⇒ *была сломана*). (пример на базе Маслов 1983)

Пример (6) показывает, однако, что форма статального перфекта может употребляться в гипотаксическом контексте, так что состояние, обозначенное глаголом в позиции первого звена цепочки, не обязательно ориентируется на момент речи, форма уместна и в таком контексте, где есть только момент наблюдения:

(6) а. *Он увидел телегу и понял, что она *сломана* и после починена. (пример из Маслов 1983)

б. Он увидел телегу и понял, что она *была сломана* и после починена.

Итак, примеры (4), (5) показывают, что форма на *н*/*т* является в речевом режиме статальным перфектом наст. времени, а пример (6) – что она не является ПЕРВИЧНЫМ эгоцентриком, т.е. формой, свойственной исключительно речевому режиму: она возможна и в гипотаксическом контексте, где обозначает состояние, актуальное на момент наблюдения. Так что статальный перфект, как и сов. вид, отличается от англ. формы Present Perfect, которая невозможна в гипотаксическом контексте.

В отличие от формы сов. вида, форма статального перфекта сочетается с обстоятельством прош. времени, характеризующим время события, **не теряя перфектности**, т.е. идеи актуальности состояния в момент речи. Это доказывает пример (7):

(7) Церковь Спаса на Нередице *разрушена* во время второй мировой войны.

В настоящий момент состояние перестало быть актуальным, и предложение понимается как ложное.

Посмотрим теперь, как ведет себя эта форма в нарративном режиме (использован материал из Князев 1989). В нарративном тексте, в котором базовое время прошедшее, статальный перфект недопустим, поскольку форма без связки трактуется как наст. время. Поэтому в нарративном контексте в примерах (196), (197)[1], где базовое время прош., употребляется форма со связкой:

(196) <Я предложил нечто.> Но мое предложение *было выслушано* с недоумением, а потом меня просто выгнали

[1] Нумерация примеров – по Князев 1989.

064-Paducheva:_Layout 1 18.05.2008 16:02 Page 416

на улицу (Шефнер: 372);

(197) <...> Очнулся в кустах, <...> дополз с трудом до шоссе, где и *был подобран* (Богомолов: 151).

Статальный перфект в этом контексте невозможен:

(201′) белые *искали* меня и не могли найти. *Вместо меня *арестован* гимназист К.

В противоположность этому, в (199), где имитируется речевой режим и возможно наст. речевое, возможна и статально-перфектная форма:

(199) Сегодня, 16 августа, <...> *уничтожена* остаточная группа противника (Богомолов: 74).

Статальный перфект возможен и в контексте с базовым наст. историческим, т.е. там, где режим интерпретации нарративный, см. (201′′):

(201′′) белые ищут меня и не могут найти. Вместо меня *арестован* гимназист К.

В Князев 1989 нарративным называется то употребление глагольных форм, которое мы назвали цепочечным.[2] Ключевым для Ю.П.Князева является пример (201), в котором невозможна цепочка статально-перфектных форм:

(201) белые *искали* меня <...> Вместо меня *был арестован* и *отпущен* гимназист К. (Каверин: 273).

Наш анализ показывает, однако, что в контексте примера (201), с базовым прош. временем нарратива, невозможна не только цепочка (о которой автор говорит, что она имеет нарративную, или повествовательную функцию), но и единичный статальный пассив, см. (201′). В то же время в (201′′′), где базовое время настоящее, возможен не только единичный пассив, но и цепочка:

(201′′′) белые *ищут* меня и не могут найти. Вместо меня *арестован* и *отпущен* гимназист К.

Получается, что цепочка статально-перфектных форм, невозможная в речевом режиме, пример (5), возможна в нарративном в контексте базового настоящего.

Пример 4. Персональная интерпретация утвердительной модальности. Открытием в теории нарратива можно считать обнаружение того факта, что персональную интерпретацию в НПР может иметь **утвердительная модальность** (индикативное наклонение). Сам факт наличия субъекта (всегда подразумеваемого) у индикатива не был до сих пор предметом внимания. То, что субъект есть, скажем, у оптатива, не вызывает сомнений – естественно, что у желания имеется субъект[3]. Нельзя, однако, не признать, что и за утвердительной модальностью тоже стоит субъект сознания.

Как известно, в речевом режиме на утверждающем лежит ответственность за его утверждение – он несет так называемое эпистемическое обязательство (Парадокс Мура). В традиционном нарративе субъектом эпистемического обязательства, которое входит в семантику индикатива, всегда является говорящий, так что индикатив – первичный эгоцентрик. Между тем в НПР индикатив может оказаться, как и другие первичные эгоцентрики, в распоряжении 3 лица. Пример (из Tammi 2003):

(1) He [Frank Churchill] stopped again, rose again, and seemed quite embarrassed. – He was more in love with her than Emma had supposed <...>. (Джейн Остин. Эмма)

Второе предложение этого текста передает впечатление Эммы – которое, как читатель скоро узнает, не соответствует действительности. Использование НПР – это способ, которым автор пользуется для того, чтобы ввести читателя в заблуждение: субъектом индикатива в данном контексте является не повествователь, а персонаж – который имеет право ошибаться.

3. Слова

Пример 1: *сейчас* 4 (Мельчук 1985) как *сейчас* 1 нарративного режима. На примере значений слова *сейчас* как нельзя лучше демонстрируется зависимость значения слова от режима интерпретации.

Словари русского языка (в частности, Словарь языка Пушкина) различают у *сейчас* три значения:

сейчас-1 = ‘в настоящий момент’ (с настоящим временем глагола);

сейчас-2 = ‘только что’ (с глаголом в прош. времени);

сейчас-3 = ‘в ближайшем будущем’ (с глаголом в буд. времени).

В канонической коммуникативной ситуации, т. е. при дейктической интерпретации, слово *сейчас* может употребляться во всех трех своих значениях:

(1) Я *сейчас* пишу письмо (*сейчас*-1);

(2) Я *сейчас* писал письмо (*сейчас*-2);

[2] Скорее всего, нарративным это значение называют исходя из формулировки (получившей широкое распространение), согласно которой в нарративах соблюдается Принцип хронологического порядка (изложения событий) – соблюдение этого принципа иногда принимается за само определение нарратива.

[3] См. в Плунгян 2000: 317 о том, что грамматические показатели модальности описывают, в разных языках, либо точку зрения субъекта, либо говорящего.

064-Paducheva:_Layout 1 18.05.2008 16:02 Page 417

(3) Я *сейчас* буду писать письмо (*сейчас*-3).

Сейчас-1 выражает в речевом режиме одновременность ситуации с моментом речи, т. е. с настоящим временем говорящего. Отсюда аномалия в (4) (пример из Мельчук 1985):

(4) К Мише нас не пустили. *Он был *сейчас* с дамой.

В самом деле, при интерпретации в речевом режиме возникает противоречие: *сейчас*-1 выражает одновременность ситуации с моментом речи, а прошедшее время глагола (*был*) представляет ситуацию как предшествующую этому моменту. Нужно было употребить адвербиал *в тот момент*, который служит нарративным эквивалентом для *сейчас* 2.

В примере (5), где *сейчас* совместимо с прош. временем глагола, И.А.Мельчук толкует *сейчас* как *сейчас* 4 = ‘в данный момент, имевший место в прошлом’:

(5) *Сейчас* он внушал жалость.

Между тем, *сейчас* 4 – это просто *сейчас* 1 в нарративном контексте (единственное из трех значений *сейчас*, которое является вторичным эгоцентриком и имеет одинаковую интерпретацию в обоих режимах).

Пример 2. Местоименные наречия *здесь* и *там* различаются примерно так же, как *сейчас* 1 и *сейчас* 2. А именно, *здесь* – вторичный эгоцентрик и употребляется в нарративе практически не меняя значения; а *там*, которое в речевом режиме каком-то смысле диалогично, т.е. принимает во внимание не только говорящего, но и адресата, в нарративе меняет значение: утрачивает семантику отдаленности и употребляется как чистый анафор.

Пример 3. Вводное *оказывается* предполагает несколько субъектов – Источник сведения, получатель сведения-знания и субъект удивления:

(1) Нашлись, нашлись! Они, *оказывается*, болели и не подавали весточек! (Л. Петрушевская)

В (2) это разные лица: Источник сведения – собеседница-квартирохозяйка, а главный персонаж с удивлением передает состояние сознания собеседницы, которое не становится его знанием (пример Г.С. Храковского, с другой интерпретацией):

(2) – Я вам скажу. Хотите откровенно? Я давно замечаю за Вами, Дима. – И тут она понесла такой немыслимый и ошеломляющий вздор, что Глебов онемел от изумления. *Оказывается*, он с каким-то особенным вниманием всегда осматривает их квартиру, на кухне его интересовали холодильник под окном и дверь грузового лифта. Однажды он подробно расспрашивал <...> (Ю.Трифонов)

В примере (3), речевой режим, это прямая речь:

(3) Я, *оказывается*, люблю другую женщину (В.Набоков. Машенька).

В тексте Набокова говорящий (Ганин) употребляет эту странную фразу в ходе небрежного объяснения с надоевшей любовницей. Здесь словом *оказывается* описано получение говорящим сведений о самом себе. Т.е. Ганин является, одновременно, Источником сведения, субъектом знания и субъектом удивления – что и объясняет ее странность.

4. Конструкции

Дейктическая конструкция нам известна только одна. Это конструкция с генитивом отрицания в контексте глагола *быть*. В Падучева 1992 семантическое различие между генитивной конструкцией в (1а) и номинативной в (1б) было объяснено с обращением к фигуре Наблюдателя. Фраза (1а) уместна в устах человека, который находится (или имеет своего представителя) в школе. А (1б) не предполагает Наблюдателя:

(1) а. Коли нет дома (ср. *Коли нет в Лондоне*);

б. Коля не дома.

Генитив отрицания при глаголе *быть* выражает наблюдаемое отсутствие.

В речевом режиме Наблюдателем является говорящий. Поэтому для фразы (2), где субъект – говорящий, в ситуации, когда говорящий дома, прямая интерпретация дает противоречие: говорящий должен быть дома, чтобы наблюдать свое отсутствие. Единственное возможное понимание – в значении своего рода несобственной прямой речи. Обычно говорящий хочет, чтобы про него так отвечали по телефону:

(2) °Меня нет дома.

Естественно, что в (3), в нарративном режиме, аномалия пропадает – Наблюдатель не говорящий, а врач (изменение режима меняет ориентацию Наблюдателя):

(3) Приходит врач, а меня нет дома. [наст. нарративное]

В примере (4) говорящий, т.е. лицо, обозначенное местоимением я, не Наблюдатель, а Субъект сознания, который **мыслит себя** в некотором месте даже тогда, когда его там нет, так что фраза не аномальна и при интерпретации в речевом режиме:

(4) Меня не было в Москве.

064-Paducheva:_Layout 1 18.05.2008 16:02 Page 418

В самом деле, (4) звучит, в контексте разговора двух москвичей, гораздо естественнее, чем, скажем, (5):
(5) Меня не было в Лондоне.

Между тем в примере (6) генитив неуместен (контекст: женщина стоит в очереди в сбербанк; у нее звонит мобильник; она отвечает клиенту, объясняя, почему она в этот момент не может дать ему нужной справки):
(6) *Меня нет в офисе (надо сказать – *Я не в офисе*).

В этой ситуации существенно, где человек реально находится, а не где он себя мыслит. Сопоставление примера (6) с примером (4) показывает, что важно различить место в котором говорящий находится, и место, которое он мыслит как свое (см. понятие личной сферы говорящего в Апресян 1986). Так что для субъекта 1 лица генитив при локативном *быть* в наст. времени полностью исключен.

Отсутствие человека в месте, которое он мыслит как свое, маркируется просодически (Падучева 2004: 460); так, в (7а) ремой является только *нет*, а в (7б) – вся группа сказуемого в целом:

(7) а. Иванова нет \ в Лондоне;
б. Иванова нет в Москве \ .

———

Приведенные примеры показывают, что режим интерпретации – важный фактор, определяющий динамическую семантику эгоцентрического элемента. Следует подчеркнуть, что речевой режим – отнюдь не то же, что разговорная речь. Скажем, слово *отныне* – вторичный эгоцентрик и вполне допускает интерпретацию в речевом режиме, хотя относится к высокому стилю.

Список литературы

1. Апресян Ю.Д. Дейксис в лексике и грамматике и наивная модель мира // Семиотика и информатика, вып. 28, М.: ВИНИТИ, 1986, с. 5-33.
2. Виноградов В.В. Избранные труды. Поэтика русской литературы. М., 1976.
3. Князев Ю. П. Акциональность и статальность: их соотношение в русских конструкциях с причастиями на -н, -т. München: Otto Sagner, 1989. (Specimena philologiae slavicae, Bd 81).
4. Мельчук И.А. Семантические этюды. I. «Сейчас» и «теперь» в русском языке.- «Russian linguistics», vol. 9, Nos. 2-3, 1985.
5. Падучева Е. В. Семантика вида и точка отсчета // Изв. АН СССР. Сер. лит. и яз. 1986. Т. 45. № 5. С. 413–424.
6. Падучева Е.В. Семантические исследования. Семантика времени и вида в русском языке. Семантика нарратива. М.: Языки русской культуры, 1996.
7. Плунгян В. А. Общая морфология. М.: УРСС, 2000.
8. Успенский Б.А. Поэтика композиции. М.: Искусство, 1970.
9. E.Benveniste. Les relations de temps dans le verbe français. BSL, t. 54, 1959. Русский перевод: Э.Бенвенист. Общая лингвистика. М.: Прогресс, 1974.
10. Tammi P. Risky business: probing the borderlines of FID. Nabokov's An affair of honor (Podlec) as a test case. //Linguistic and literary aspects of free indirect discourse from a typological perspective. Tampere, Tamperen yliopisto taideaineiden laitos, 2003, 41-54.

065-Palko_M:_Layout 1 18.05.2008 16:03 Page 419

ИНТОНАЦИЯ НЕЗАВЕРШЕННОСТИ ТЕКСТА В НЕМЕЦКОМ ЯЗЫКЕ ВСОПОСТАВЛЕНИИ С РУССКИМ

INTONATION OF THE GERMAN COHERENT DISCOURSE IN CONTRAST TO THE RUSSIAN ONE

Палько М.Л. *(m_palko@mail.ru)*
Институт языкознания РАН

В данной работе исследуются интонационные средства выражения значения незавершенности текста в немецком языке, в частности, в контексте контраста и эмфазы.

Данная работа посвящена сопоставительному анализу интонационных средств выражения коммуникативного значения незавершенности текста, то есть указания на то, что текущее предложение не последнее и что продолжение повествования следует. Интонация понимается как средство формирования предложения, создания его целостности, выделения в нем коммуникативных компонентов и указания на его место в структуре текста. Интонация складывается из сочетания движения тона, силы звука, тембра, длительности и других признаков. Целью данной работы является исследование интонационных показателей незавершенности текста.

При изучении интонационных средств выражения текстовой незавершенности в немецком языке мы опираемся на системное описание акцентных средств демонстрации незавершенности в русском языке, представленное в работах Т.Е.Янко [5-9]. Интонационные средства выражения коммуникативного значения незавершенности текста в немецком языке ранее не описывались, детально исследовались только фразовая интонация и мелодические типы предложений, а исследования по интонации текста были преимущественно посвящены принципам распределения пауз в предложении и в тексте в [2;10;14].

Материалом нашего исследования послужили звучащие тексты – фрагменты передач немецкого радио Deutsche Welle, телеканала Euronews, фрагменты произведений классической немецкой литературы, прочитанные информантами-носителями языка, и неподготовленные рассказы информантов. Аудиоданные радио- и телепередач, репортажей, фрагментов ток-шоу, новостных сообщений и рассказов 10 информантов-носителей немецкого языка, мужчин в возрасте от 20 до 26 лет, общим звучанием приблизительно 100 часов составили корпус экспериментального материала. Коллекция звучащих текстов хранится в электронном виде.

Методика, которой мы следуем в нашем исследовании при работе с аудиоданными, разработана Н.Д.Светозаровой [4]. Алгоритм работы с аудиоданными включает в себя несколько этапов. Первый этап – подготовительный – это отбор материала. Второй этап – слуховой и инструментальный анализ. Главным при анализе речи является слуховой анализ, то есть при прослушивании фрагмента речи коммуникативно релевантный пик и слово носитель этого пика определяются на слух, а инструментальный анализ используется для проверки гипотез и для передачи на бумаге параметров звучащей речи. Для инструментального анализа мы используем компьютерную систему анализа звучащей речи Speech Analyzer. При анализе примеров ниже используются тонограммы, которые отражают изменение частоты тона в герцах, а при анализе примера (9) используется также график интенсивности. Заключительный этап – интерпретация полученных данных, в ходе которой делаются выводы о значимости тех или иных просодических характеристик.

Перейдем к изложению результатов, полученных в связи с анализом интонационных показателей незавершенности текста.

Мы предполагаем, что в немецком языке для указания на незавершенность текста используются, как минимум, четыре типа акцентов. Под акцентом мы понимаем основную единицу интонации, которая является носителем некоторых значений, фиксируется на словоформе-акцентоносителе и обладает тонально-темпоральными характеристиками, такими как движение тона на ударном слоге акцентоносителя и заударных слогах, если они есть, растяжки, аллегро и др. признаки. Акценты складываются в интонационные контуры предложений. В немецком языке для указания на незавершенность текста используются следующие типы акцентов. Это восходящее движение тона на ударном слоге типа русской интонационной конструкции ИК-3 по Е.А.Брызгуновой, нисходяще-восходящее движение тона типа ИК-4, восходящее движение тона на ударном слоге плюс долгий ровный высокий тон на заударных слогах типа ИК-6 по Е.А.Брызгуновой, а также особое, отсутствующее у Е.А.Брызгуновой, восходящее движение тона на ударном слоге с последующим подъемом на

065-Palko_M:_Layout 1 18.05.2008 16:03 Page 420

заударных слогах. Таким образом, при описании немецких и русских примеров мы существенно опираемся на систему интонационных конструкций русского языка, разработанную Е.А.Брызгуновой [3, 97-111]. Перейдем к рассмотрению примеров.

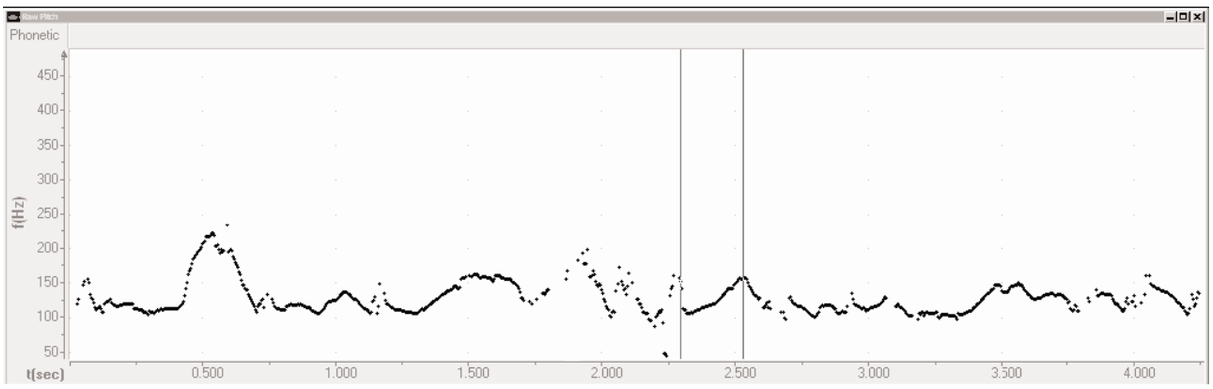
1. Немаркированные средства выражения незавершенности текста
1.1. Стратегия с использованием акцента типа ИК-3

Наиболее частотным средством указания на незавершенность текста в немецком языке, как и во многих языках, служит акцент, который характеризуется подъемом частоты основного тона на ударном слоге акцентоносителя незавершенности плюс падение тона на заударных слогах, если они есть.

Рассмотрим пример (1).

(1) *Immerhin werden da vom Münchener Infoinstitut 7000 Unternehmen befragt, wie sie denn ihre Lage einschätzen, und vor allem wie sie nach vorne schauen, < und da haben die allermeisten gesagt: «Wir schauen sehr optimistisch in die Zukunft»>*.

'Тем не менее, в Мюнхенском информационном институте было опрошено 7000 предприятий на предмет того, какони оценивают свое положение, и, прежде всего, как они смотрят в будущее, <и большинство ответило: «Мы смотримв будущее с большим оптимизмом». >'



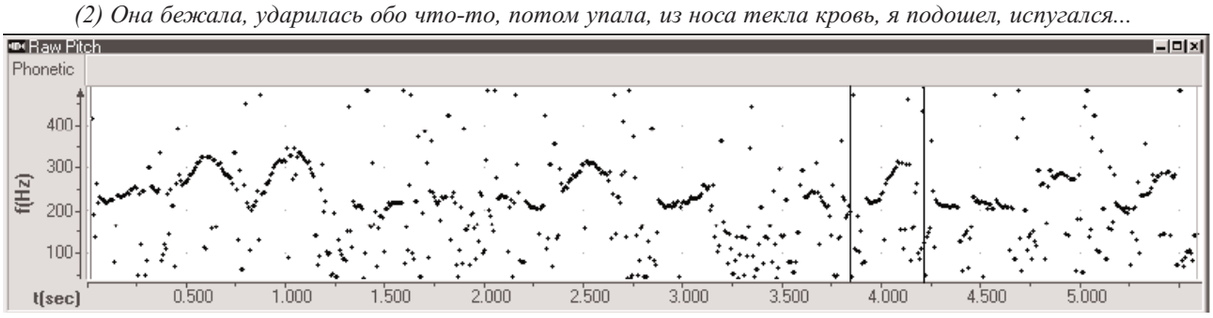
.... *Sieben*	*Tausend*	*Unternehmen*		*befragt,*	*wie sie denn ihre Lage einschätzen…*
… Семь	тысяч	предприятий		опрошено	как они же свое положение оценивают…

Тонограмма 1

В данном примере незавершенность текста выражается подъемом тона на словоформе *befragt*. На тонограмме 1 курсорами выделен ударный слог словоформы. Такое движение тона с подъемом на ударном слоге достаточно близко ИК-3 по Е.А.Брызгуновой.

В русском языке аналогичная стратегия незавершенности является одним из самых частотных – немаркированных – способов указания на то, что продолжение повествования впереди. Рассмотрим русский пример и соответствующую ему тонограмму. Этот пример заимствован из работы [7] и является фрагментом рассказа ребенка.

(2) Она бежала, ударилась обо что-то, потом упала, из носа текла кровь, я подошел, испугался...



Она бежала, ударилась обо что-то, потом упала из носа текла кровь, я подошел, испугался...

Тонограмма 2

065-Palko_M:_Layout 1 18.05.2008 16:03 Page 421

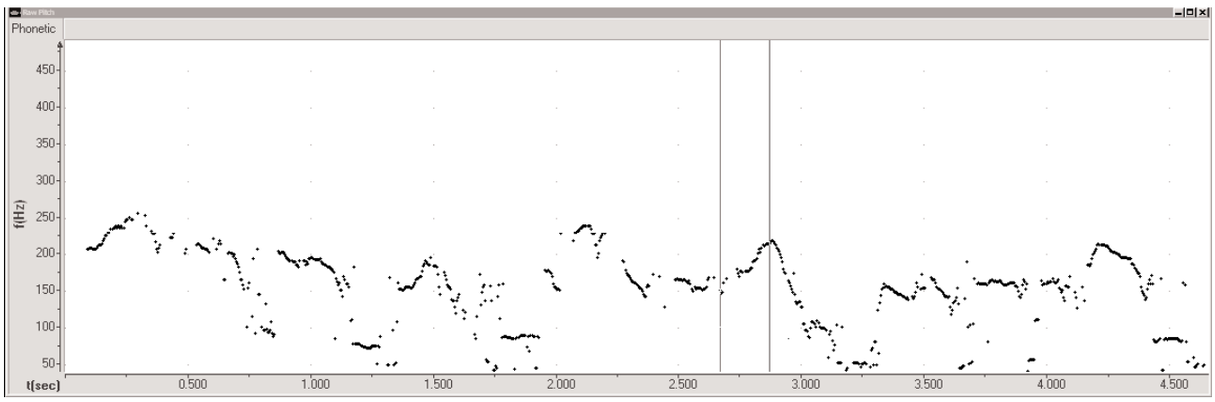
Тонограмма показывает, что на словоформах *бежала, ударилась, упала, кровь, подошел* и *испугался* фиксируется акцент ИК-3 с характерным для него подъемом на ударном слоге и падением на заударных, если они есть. В данном примере словоформа кровь (выделена курсорами) в отсутствие заударных слогов несет на себе фактически такое же движение тона, что и словоформа *befragt* в немецком примере (1).

Пример (3) ниже демонстрирует, что данная стратегия поддержания связности текста может комбинироваться со значением эмфазы. Значение эмфазы мы понимаем как выражение сильных чувств говорящего по поводу ненормативных явлений жизни. В работах [6; 8] было показано, что эмфаза может комбинироваться с темой, ремой, вопросом и императивом. Однако материал примера (3) говорит о том, что эмфаза вступает в композиции не только с темой, ремой и вопросом, но также и с незавершенностью нарратива. Ранее этот факт, насколько нам известно, нигде не отмечался. Эмфаза характеризуется увеличением долготы ударного слога в среднем в полтора раза, по сравнению с обычным произнесением, существенным удлинением согласных, тональным перепадом, состоящим в присоединении к образующему – восходящему или нисходящему – тону отклонения движения тона в противоположную сторону [8, 64].

Обратимся к примеру (3). Это фрагмент интервью, в котором режиссер документального фильма рассказывает о целях манифестаций, которые организуют молодежные политические партии. Он подчеркивает, что современное молодое поколение сильно отличается от поколения своих родителей, у молодежи другие идеалы и принципы.

(3) Die junge Generation ist anders, sie fühlt sich europäisch, strebt nach Demokratie.

‘Молодое поколение другое, оно чувствует себя по-европейски, стремится к демократии’.



Die junge Generation ist	*anders,*	*sie fühlt sich europäisch, strebt nach Demokratie*
Молодое поколение	другое,	оно чувствует себе по-европейски, стремится к демократии

Тонограмма 3

В данном примере значение незавершенности выражено восходящим движением тона на ударном слоге словоформы *europäisch* (она выделена курсорами на тонограмме). На выделенном фрагменте наблюдается подъем тона на ударном слоге *-pä-* и предшествующее подъему небольшое отклонение тона в противоположную сторону. Такое движение тона маркирует значение эмфазы. Ударный слог словоформы-акцентоносителя характеризуется также повышенной длительностью. Аналогичные интонационные показатели наблюдаются также на словоформах *anders* и *Demokratie*. Таким образом, в данном примере значение незавершенности сочетается со значением эмфазы.

1.2. Стратегия с использованием акцента типа ИК-4

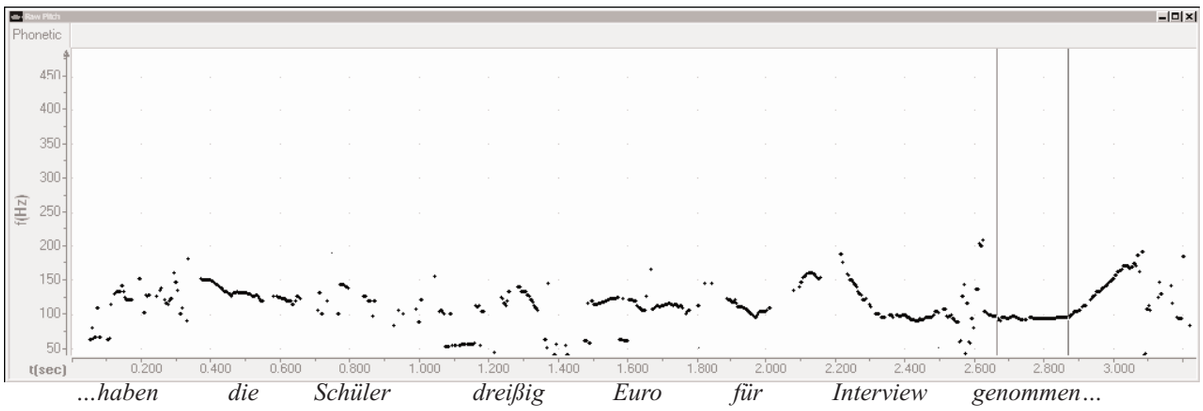
Указание на незавершенность повествования с помощью нисходяще-восходящего движения тона на акцентоносителенезавершенности также весьма частотная модель немецкой речи.

Рассмотрим пример (4). Это фрагмент рассказа радиожурналиста о съемке репортажа в школе.

(4) <Am zweiten Tag, als immer mehr Kamerateams an diese einzelne Schule kam>, haben die Schüler 30 Euro für Interview genommen < und, sozusagen, selber sich für die Medien inszeniert.>

‘<На второй день, когда пришло еще больше репортеров>, школьники получили по 30 евро за интервью <и, так сказать, сами подыгрывали им.>’

065-Palko_M:_Layout 1 18.05.2008 16:03 Page 422

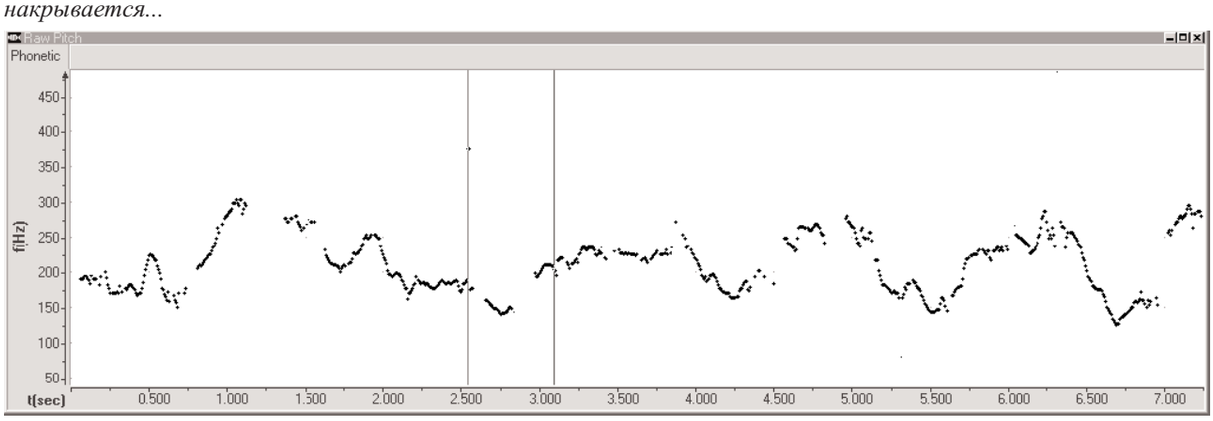


Тонограмма 4

Тонограмма 4 показывает, что на ударном слоге словоформы genommen фиксируется ровный низкий тон (он выделен курсорами), а на заударный слог приходится восходящий фрагмент акцента. Таким образом, в данном примере акцент типа ИК-4 реализуется на словоформе, имеющей заударные слоги.

В русском языке стратегия с ИК-4 в качестве показателя незавершенности менее частотна, чем стратегия с ИК-3. Кроме того, в русском языке ИК-4 придает тексту специальное значение «рассказа по порядку». Говорящий заранее выделяет множество объектов или событий, о которых он намерен говорить, и выстраивает их в цепочку. Такой тип акцента в русском языке характерен для объяснений на уроках и лекциях, в официальном стиле речи при стремлении говорящего подчеркнуть последовательность в изложении, в бытовом общении при изложении событий «по порядку», например, при изложении кулинарных и иных рецептов [6]. Обратимся к русскому примеру (5). В этом примере гадалка рассказывает о способе гадания на зеркале. Пример взят из работы [6].

(5) Значит, зеркало, чистое новое полотенце обязательно, тарелочку большую, поднос большой, накрывается...



Тонограмма 5

Акцент ИК-4 фиксируется на словоформах зеркало, полотенце, тарелочку, поднос (большой) и накрывается. На тонограмме курсорами выделены ударные и заударные слоги словоформы полотенце. На ударный слог словоформы полотенце приходится падение тона, на заударный слог – подъем.

Итак, сравнительный анализ русского и немецкого материала показывает, что нисходяще-восходящее движение тона на акцентоносителе незавершенности в русском языке служит выразителем стратегии т.н. «рассказа по порядку», а в немецком – это немаркированный – нейтральный – показатель незавершенности.

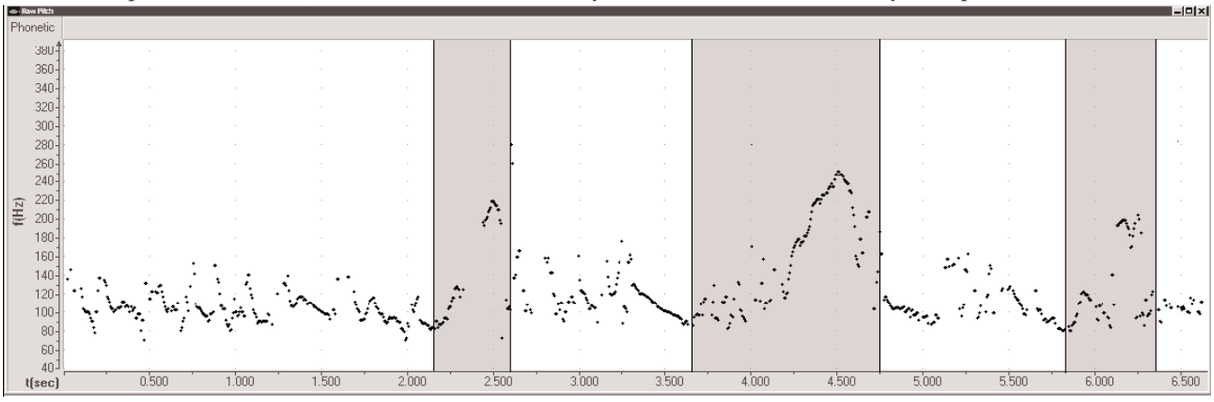
065-Palko_M:_Layout 1 18.05.2008 16:03 Page 423

2. Незавершенность в контексте «рассказа по порядку»

Материал примера (6), который анализируется ниже, и других говорит о том, что в немецкой речи для выражения значения «рассказа по порядку» используется стратегия с подъемом тона на ударном слоге с дальнейшим подъемом тона на заударных слогах словоформы-акцентоносителя незавершенности. Такое движение тона в системе интонационных конструкций Е.А.Брызгуновой отсутствует. На русском материале оно впервые было обнаружено Т.Е.Янко [6]. Рассмотрим пример (6) и соответствующую ему тонограмму. В данном примере речь идет о частной коллекции картин, которая будет выставлена в разных картинных галереях мира. Организатор выставок рассказывает о своих планах на год. Он ясно представляет себе, что он последовательно будет делать каждый месяц, и подробно рассказывает об этом.

(6) Wir werden erzählen, was wir dieses Jahr machen. Wir beginnen mit Sizilien schon in einer Woche…

‘Мы расскажем о том, что мы делаем в этом году. Мы начинаем с Сицилии уже через неделю…’



Wir werden erzählen, was wir dieses Jahr machen.	*Wir beginnen*	*mit*	*Sizilien*	*schon in einer Woche…*
Мы расскажем о том, что мы делаем в этом году.	Мы начинаем	с	Сицилии	уже через неделю…

Тонограмма 6

Тонограмма показывает, что на словоформах *machen, Sizilien* и *Woche* фиксируется одинаковое движение тона. Так, на ударном слоге словоформы machen наблюдается восходящее движение тона, на заударном слоге подъем тона не прекращается и продолжается вплоть до конца словоформы. Возникает гипотеза о том, что в немецком языке это движение тона маркирует «рассказ по порядку». Что касается аналогичного акцента в русском языке, в работе [6] было показано, что такое движение тона в русской речи тоже моделирует изложение череды событий, но это не обязательно рассказ по порядку. В русском языке этот акцент характерен для взволнованной речи. В немецком языке, в отличие от русского, подъем на ударных плюс подъем на заударных характерен не только для отличающейся повышенной эмоциональностью, но и для обычной речи. В немецкой речи этот акцент гораздо более частотен, чем в русской.

В следующем разделе будет рассмотрена незавершенность в контексте отражения мыслительной деятельности говорящего: воспоминания, размышления, ментального поиска.

3. Незавершенность в контексте воспоминания

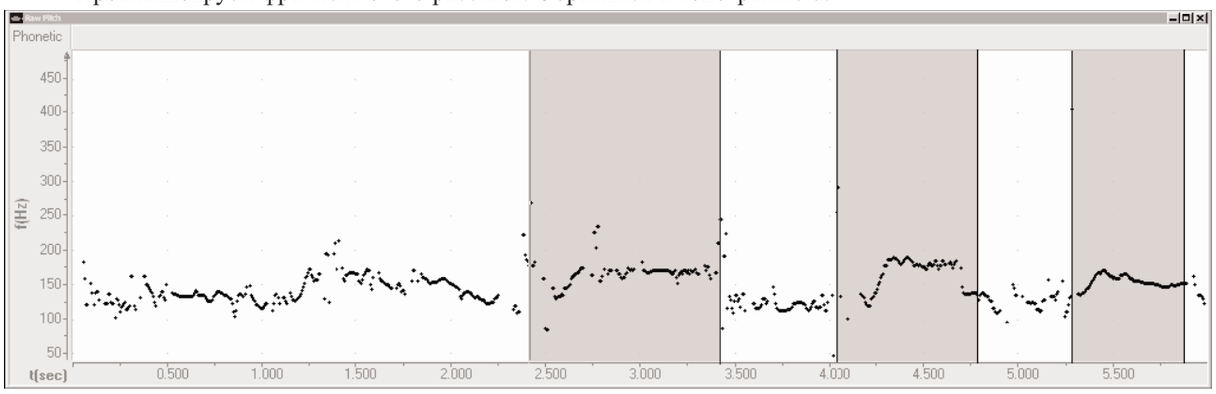
В данном разделе будет рассмотрена не незавершенность в чистом виде, а скорее, стратегия построения текста, имититрующего ментальную деятельность говорящего: рассуждения, недоумения, воспоминания, размышления о будущем, см. [6]. Это значение часто используется при изложении череды событий, т.е. фигурирует и в контексте незавершенности, когда говорящий вспоминает или размышляет о череде событий. В примере (7) ниже оно связано с отражением процесса обоснования. Речь идет о музыкальном фестивале в Лейпциге. Говорящий как бы рассуждает вслух, почему, по его мнению, музыкальный фестиваль в Лейпциге может быть интересен зрителям и исполнителям.

(7) Ich denke, das ist, was bei den Autoren auch gut einkommt, daß wir eben nicht nur große Oper spielen, sondern auch Kammeroper, sei es Lyrik oder sei es Prosa, <sei es der bekannte Star oder sei es der Debütant. Alle finden eine von den Präsentationen, die ihnen angemessen ist. Und ich glaube, das ist das Geheimnis von Leipzig.>

065-Palko_M:_Layout 1 18.05.2008 16:03 Page 424

‘Я думаю, это то, что также привлекает авторов, что мы как раз играем не только большие оперы, но и камерные концерты, будь то лирика или проза, <известная звезда или дебютант, каждый ищет достойного выражения. И я думаю, что это секрет Лейпцига.>’

Проанализируем фрагмент этого рассказа. Обратимся к тонограмме 7.



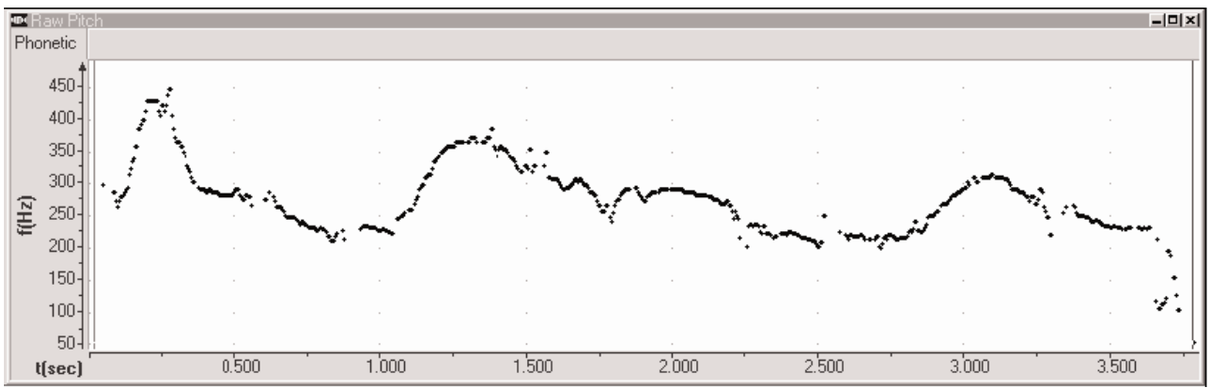
… daß wir eben nicht nur große Oper spielen, sondern auch Kammero-oper sei es Lyrik oder sei es Prosa…
…что мы как раз не только большие оперы играем, но также камерные оперы, будь то лирика или будь то проза…

Тонограмма 7

Незавершенность текста в данном примере выражается специфическим восходящим тоном на словоформах Kammeroper, Lyrik, Prosa (выделено на тонограмме). Ударный слог словоформы Kammeroper характеризуется восходящим тоном и повышенной длительностью. На заударных слогах наблюдается долгий ровный высокий тон, что отчетливо видно на тонограмме. Заударные слоги сильно растянуты. На словоформах Lyrik, Prosa, как и на словоформе Kammeroper, тоже наблюдается подъем тона на ударном слоге и высокий ровный, слегка нисходящий на словоформе Prosa, тон на заударных слогах.

Подобное выражение незавершенности текста широко используется и в русской разговорной речи, ср. пример из детской речи (пример заимствован из [9]):

(8) Мы заехали на полянку на березовую, погуляли там.



Мы заехали на поля-янку На березовую, погуля-яли там.

Тонограмма 8

Тонограмма показывает, что в словоформах *полянку* и *погуляли* предударные слоги ровные и низкие. Затем наблюдается существенный подъем на ударном слоге акцентоносителя. Ударный слог и последующие заударные – достаточно ровные и высокие, имеют продленное время звучания по сравнению с другими слогами. Это ИК-6 по Е.А.Брызгуновой.

Стратегия незавершенности, характеризующаяся акцентом типа ИК-6, может вступать в композицию с коммуникативным значением контраста. Это явление иллюстрирует пример (9). Под контрастом мы понимаем

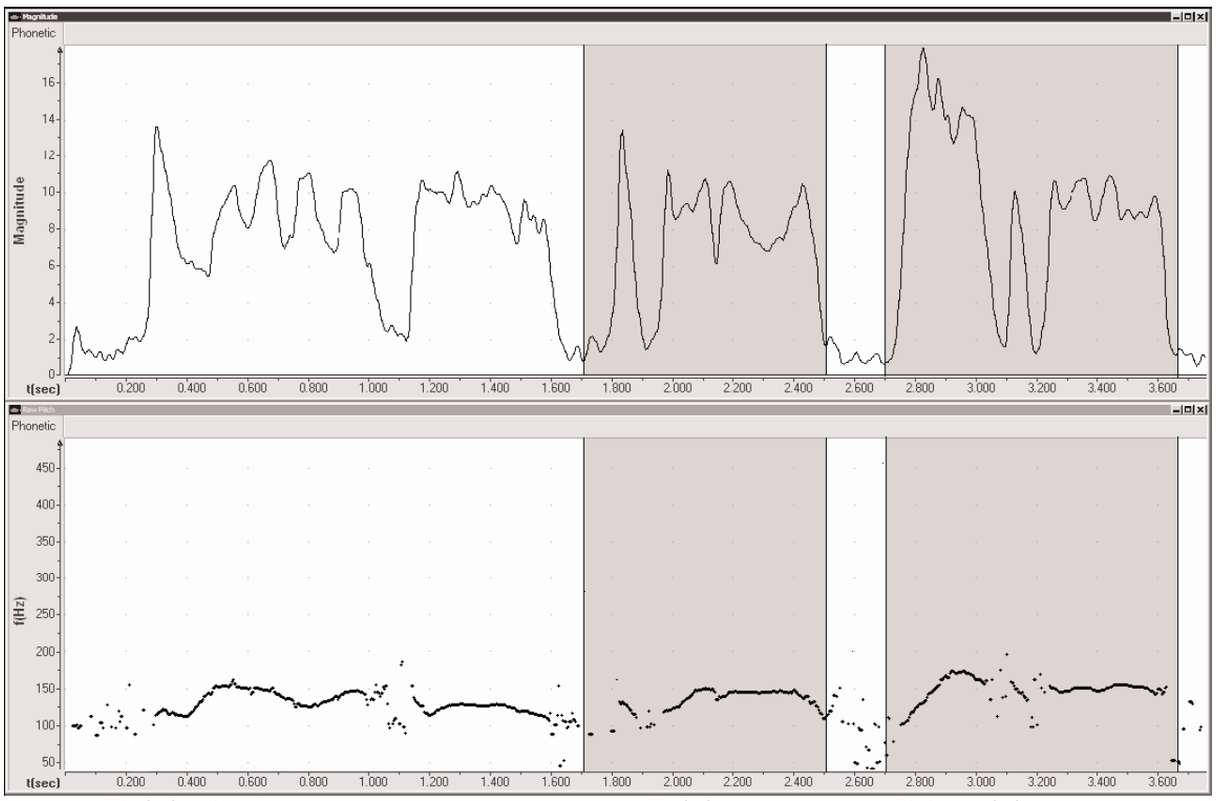
065-Palko_M:_Layout 1 18.05.2008 16:03 Page 425

соотнесение выделенного элемента с известным множеством, из которого делается выбор [8, 47] (о контрасте см. также [6, 198-205] и цитированную там литературу).

Пример (9) – это фрагмент телерепортажа о взаимодействии военной и гражданской авиации в Европе. Сотрудник авиакомпании рассказывает о том, что в настоящее время средствами военной авиации совершаются рейсовые пассажирские перевозки в разные города Германии. Говорящий говорит о высоком уровне военных самолетов, о современных требованиях к авиаперелетам. Данное предложение не последнее в рассказе.

(9) Wir haben im Moment Verkehre, Linienverkehre innerdeutsche nach München und nach Köln.

‘На данный момент у нас есть авиаперевозки, регулярные авиаперевозки внутри Германии в Мюнхен и Кёльн.’



Wir haben im Moment *a-a* *Verkehre,* *Linienverkehre …*
У нас есть на данный момент э-э авиаперевозки, регулярные авиаперевозки…

Тонограмма 9

Незавершенность в данном примере выражена специфическим движением тона на словоформах *Verkehre и Linienverkehre* (они выделены на графиках серым цветом). В словоформе *Verkehre* на ударном слоге *-keh-* наблюдается существенный подъем, который также характеризуется увеличенной длительностью. Высота тона сохраняется и на заударном слоге *-re*. Заударный слог характеризуется высоким ровным движением тона. Между тем данный пример имеет и интонационную специфику, характерную для контраста. В чем она выражается? В словоформе *Linienverkehre* на слоге *li-* подъем тона совершается в больших диапазонах частот и сопровождается высокой интенсивностью (изменение интенсивности тона демонстрирует верхний график). Увеличение диапазона частот и повышенная интенсивность свидетельствуют о контрасте. Контраст говорит о том, что говорящий, сказав *Verkehre* ‘перевозки’, решает внести уточнение в свою речь: не просто перевозки, а *Linienverkehre* ‘регулярные перевозки’. Вторая часть сложного слова характеризуется, как и в первом случае, подъемом тона на ударном слоге *-keh-* и сильно растянутой, ровной заударной частью, что указывает на незавершенность. Таким образом, пример (9) демонстрирует композицию коммуникативных значений ‘незавершенность текста’ плюс ‘контраст’.

065-Palko_M:_Layout 1 18.05.2008 16:03 Page 426

***

Итак, анализ текстов позволил выделить в немецком языке акценты, указывающие на незавершенность текста. Незавершенность может комбинироваться со значениями контраста и эмфазы. Было показано, что немецкий язык использует, как минимум, четыре акцентных средства демонстрации незавершенности. Акцент, близкий русской интонационной конструкции ИК-3, по Е.А.Брызгуновой. Он характеризуется подъемом тона на ударном слоге и падением на заударных слогах, если они есть, и служит немаркированным показателем незавершенности текста, как в русском, так и в немецком языках. Нисходяще-восходящий акцент типа ИК-4 в русском языке говорит об особом типе незавершенности – «рассказе по порядку», а в немецком языке подобное значение у нисходяще-восходящего акцента отсутствует и он, как и восходящий акцент типа ИК-3, служит немаркированным показателем незавершенности. Средством выражения так называемого «рассказа по порядку» в немецком языке служит акцент с подъемом тона на ударном и последующим подъемом на заударных слогах. Акцент, близкий ИК-6, характеризующийся восходящим, долгим тоном на ударном слоге и высокой, долгой, ровной заударной частью, связан с имитацией мыслительной деятельности: рассуждения, воспоминания, мечтания, резонирования, в частности, в контексте размышлений о череде событий. И наконец, различные типы незавершенности, как в русском, так и в немецком языках, способны компоноваться с коммуникативными значениями контраста и эмфазы.

Список литературы

1. Брызгунова Е.А. Практическая фонетика и интонация русского языка. М., 1963.
2. Кравченко М.Г., Зыкова М.А., Светозарова Н.Д. Ударение и интонация в немецком языке. Л.,1973.
3. Русская грамматика. М., 1982. Т.1.
4. Светозарова Н.Д. Интонационная система русского языка. Л.,1982.
5. Янко Т.Е. Датская интонация в сопоставлении с русской // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции Диалог 2005 по компьютерной лингвистике и ее приложениям. Звенигород, 2005.
6. Янко Т.Е. (в печати) Интонационные стратегии русской речи в сопоставительном аспекте.
7. Янко Т.Е. Интонация связного текста // Труды международной конференции Диалог 2006. М., 2006.
8. Янко Т.Е. Коммуникативные стратегии русской речи. М.,2001.
9. Янко Т.Е. Русская интонация в задачах и примерах // Русский язык в научном освещении. 2004. №2 (8).
10. Essen O. von. Grundzüge der hochdeutschen Satzintonation. Düsseldorf. Ratingen, 1956.
11. Majer J. Intonation und Bedeutung. Aspekte der Prosodie-Semantik-Schnittstelle im Deutschen. Stuttgart, 1997.
12. Mayer J. Transcription of German Intonation. The Stuttgart System. www.ims.unistuttgart.de/phonetik/joerg/papers/STGTsystem.ps.gz
13. Pierrehumbert J.B. The Phonology and Phonetics of English Intonation. 1980.
14. Stock E., Zacharias Chr. Deutsche Satzintonation. Leipzig,1973.

066-Pereverzeva-Kreidlin:_Layout 1 18.05.2008 16:04 Page 427

ТЕЛЕСНОСТЬ И НЕКОТОРЫЕ ОСОБЕННОСТИ СЕМИОТИЧЕСКОГО ДИАЛОГИЧЕСКОГО ПОВЕДЕНИЯ[1]

CORPOREALITY AND SOME PECULIARITIES OF SEMIOTIC BEHAVIOUR IN A DIALOG

Переверзева С. И. *(P_Sveta@hotmail.com),* ***Крейдлин Г. Е.*** *(gekr@iitp.ru)*
Российский государственный гуманитарный университет

Настоящая работа продолжает серию исследований, посвящённых механизмам и способам взаимодействия невербальных и вербальных кодов в коммуникативном акте. В работе предлагается модификация некоторых синтаксических правил, определяющих такое взаимодействие. Показано, что отдельные смысловые компоненты в толковании жеста регулярно соотносятся с теми или иными компонентами формы жеста (телесными компонентами).

Ключевые слова: диалог, вербальный знак, невербальный знак, части тела, признак

1. Тело в диалоге. Признаки частей тела

Современное представление о человеческом теле и телесности включает в себя осознание того факта, что тело представлено по-разному не только в разных естественных языках, но и в разных невербальных семиотических кодах, главным из которых, безусловно, является язык жестов. Жесты здесь понимаются в широком, родовом, смысле слова. К жестам относятся знаковые движения тела и его частей – рук, ног, головы, плеч и др., а также мимика (выражения лица), позы, взгляды и комплексные знаковые формы – манеры. Механизмам и способам взаимодействия вербальных и невербальных единиц были посвящёны статьи Крейдлин 2006; 2007; 2008.

В этих статьях речь шла о дидактических, или лекторских, жестах и их соотношении с речевыми единицами: в первой из них – о жестовых ударениях, ритмизующих, упорядочивающих речевой поток и выделяющих в нём те или иные коммуникативные единицы, а в двух других – о дейктических жестах. Особое внимание обращалось на форму и смысл жестов, меньше внимания уделялось их синтаксису и прагматическим свойствам. В центре всех этих работ было тело человека и телесное поведение самой разной природы. В частности, было показано, что в русской культуре конкретные жесты обычно согласуются с определёнными речевыми высказываниями или актами речи, такими, как предупреждение, угроза, представление-знакомство, указание и т. д. Например, выражение угрозы взрослого по отношению к ребёнку хорошо соединяется с жестами пальцев или кулака и плохо – с жестами ладони. Представление-знакомство, наоборот, плохо сочетается с жестами пальцев, а предложение мириться (у детей) – хорошо (ср. жест **мириться**, осуществляемый мизинцами).

Описывая взаимодействие невербальных и вербальных единиц, удобно воспользоваться введённым в нашей статье Крейдлин, Переверзева 2008 понятием **семиотическая концептуализация тела**. Семиотическая концептуализация представляет собой естественное расширение известного в лингвистике понятия **языковой концептуализации фрагмента мира**. Понятие семиотической концептуализации тела отражает то, что и как обычный человек говорит о теле и его частях и как он использует своё тело в коммуникативном акте. Ниже мы покажем, что без знания того, что и как человек говорит о теле и пользуется телом, то есть без знания семиотической концептуализации тела невозможно с нужной степенью полноты и точности описать особенности функционирования естественного языка и языка тела в коммуникативном акте.

Чтобы это показать, нам придётся сначала кратко остановиться на том, из чего складывается семиотическая концептуализация человеческого тела. Она, по нашим представлениям, складывается из большой совокупности содержательных признаков и их значений, характеризующих разные аспекты телесности и телесного поведения человека. К ним относятся такие признаки, как характеристики внешней формы,

[1] Настоящая работа выполнена в рамках проекта при поддержке Российского гуманитарного научного фонда (грант РГНФ № 07-04-00203а).

066-Pereverzeva-Kreidlin:_Layout 1 18.05.2008 16:04 Page 428

конфигурация и внутреннее строение тела или данной части тела, то есть **морфологические и структурные аспекты** тела и его частей.

Приведём несколько примеров, которые нам уже доводилось как-то приводить. Описывая внешний облик того или иного человека, мы можем сказать, что *У неё кривые ноги* или что *Он криворукий,* но не, например, **Он кривоголовый* или **У него кривой живот*. Или: в русском языке есть сочетания *стройное тело, изящная фигура*, но плохо сказать **изящное туловище* или **тщедушный стан*.

Ещё один важный признак – это **положение** данной части тела относительно других её частей или тела в целом. Например, язык в норме находится во рту за зубами и не виден. Уши, напротив, видны, если, конечно, не спрятаны под головным убором или за волосами, и уж точно мы знаем, что они находятся по бокам головы. Однако язык говорит нам о том, что *ушки* могут быть *на макушке*, то есть здесь реальная анатомия человека отличается от наивного представления о ней.

При описании того, как происходит устный диалог, большое значение имеет то, что многие части тела (хотя и не все) могут двигаться и менять свое положение. Поэтому, характеризуя жесты, которые используются в диалоге и в которых данные части тела играют первостепенную роль, следует указывать положение этих частей тела. Когда человек закрывает лицо руками**,** чтобы отгородиться от событий или других людей, ладони одной руки или обеих рук находятся на лице и прикрывают глаза, а выражая благодарность, человек прикладывает ладонь к груди. В последнем случае выражение благодарности может сопровождаться словами благодарности.

Поскольку коммуникация – явление динамическое, для её описания существенно указание типовых **движений**, совершаемых телом или его частями, а также важнейших характеристик этих движений. Кроме того, важно знать, какие стереотипные **действия** совершает конкретная часть тела или каковы **действия** человека **над** частью тела (как своего, так и чужого). Для определения же того, как происходит невербальный контроль над телом и телесным поведением, необходимо отдельно обращать внимание на **влияние** как тех, так и других действий на собеседника или других людей, участвующих в данном акте коммуникации.

2. Признаки частей тела. Особенности взаимодействия смысловых и формальных компонентов невербальных единиц.

В книге Крейдлин 2002 речь шла о том, что синтаксис невербального поведения складывается из двух совокупностей правил: правил внутреннего синтаксиса, которые отражают связь невербальных единиц друг с другом и порядок их следования в устном диалоге, и правил внешнего синтаксиса, которые описывают возможность соединения жестов разной природы с вербальными единицами. И те и другие правила обычно формулируются комплексно, а именно, говорится о том, что такой-то и такой-то жест сочетается или не сочетается в диалоге с таким-то и таким-то вербальным или невербальным знаком. Между тем, как показывает накопленный опыт изучения коммуникативного поведения, в частности, лекторского (см. Крейдлин 2006; 2007), такое описание представляется не вполне адекватным.

Основным тезисом нашей работы является следующее положение, которое мы хотим далее обосновать. Невербальными единицами описания механизмов и способов взаимодействия знаков, участвующих в коммуникативном акте, должны быть в общем случае не жесты в целом, а те телесные компоненты, которые играют важную роль в его образовании. Например, телесные компоненты могут входить в состав активного органа, участвующего в производстве жеста. Или бывает так, что наличие определенных телесных компонентов в составе жестовой формы является непременным условием правильного употребления соответствующего жеста в диалоге. Так, для описания синтаксической комбинации «жест благодарности» + «словесное выражение благодарности» важно участие рук и их положение на груди. На то, что благодарность исходит из сердца, указывают не только жесты, но и такие высказывания, как *Сердечно вам признательна*; *Благодарю вас от всего сердца* и др. Здесь грудь и положение рук на ней (точнее, передаваемый таким телесным компонентом смысл) соединяется со смыслом приведённых высказываний.

В качестве телесных компонентов могут выступать признаки частей тела, участвующих в производстве жеста; о некоторых из таких признаков мы уже говорили выше.

Рассмотрим несколько примеров.

(1) В группе знаковых движений прерывания контакта с собеседником[2], таких, как **отвернуться**, или жестов **отвести глаза**, **опустить глаза**, **резко повернуть голову в сторону** и т. п., телесный компонент с наиболее высокой коммуникативной значимостью – это признак **переменной ориентации**, соответственно, тела,

[2] Этот класс жестов играет заметную роль в русском языке тела и русской культуре (см. об этом книгу Крейдлин 2002).
[3] Его подробное описание дано в СЯРЖ 2001.

066-Pereverzeva-Kreidlin:_Layout 1 18.05.2008 16:04 Page 429

глаз и головы. Именно смена ориентации часто сопровождает такие выражения, как *Видеть тебя не хочу!*; *Глаза бы мои на тебя не смотрели!*; *Стыдно!*; *Не смотри на меня!* и др.

(2) Жест **бить себя в грудь**[3] имеет сложную семантическую организацию. В его толкование входят несколько смысловых компонентов, соответствующих одному и тому же компоненту формы жеста – «контакт руки с грудью». Это:

(а) ‘слова, которые произносит жестикулирующий, он считает истинными’.

Данный смысловой компонент в русском языке жестов и в русском языке соотносится с сердцем, которое находится в груди. Положение рук на груди хорошо сочетается с этим смыслом, о чём говорят многие данные. Так, в русской культуре слова, чувства и поступки, идущие *от сердца*, считаются искренними и потому истинными – ср. выражения *От всего сердца за вас радуюсь*; *Она всегда поступает, как велит сердце*. То, что вырывается из груди, что трудно скрыть, тоже считается истинным.

О связи груди и сердца с истинностью и искренностью говорят также факты, относящиеся к русскому языку тела. Сопоставим жест **бить себя в грудь** с жестом **приложить руку к груди**. Жест **приложить руку к груди** является многозначным и в своих трёх значениях соотносится, соответственно, со смыслами ‘умоляю’, ‘убеждаю’, ‘благодарю’. А каждый из этих смыслов, в свою очередь, связан с компонентами ‘искренность’ или ‘истинность’. Таким образом, именно телесный компонент «контакт руки с грудью» в обоих жестах передаёт смысл, о котором мы здесь говорим.

(б) ‘жестикулирующий указывает на себя <а не на адресата>’[4].

В русской культуре это указание регулярно соотносится с частью тела «грудь», в отличие, например, от китайской культуры, где смысл ‘указание на себя’ передаётся физически указанием на нос, то есть физическая реализация этого смысла иная.

(в) ‘жестикулирующий ручается в том, что его слова истинны’.

О связи этого смыслового компонента с физическими, телесными компонентами говорят следующие факты. Во-первых, глагол *ручаться* и имя *рука* являются однокоренными словами; письменное поручительство закрепляется подписью, которая делается рукой, а устное поручительство часто скрепляется рукопожатием. Во-вторых, рассматриваемый смысловой компонент связан ещё с одним телесным компонентом, а именно, с грудью, эта связь выявляется только в результате более сложного анализа. Дело в том, что смысл ‘ручаться’ очевидным образом связан со смыслом ‘клясться’ – их связывает, по крайней мере, компонент ‘уверенность субъекта в истинности произносимых слов’. Клятва же является не только обыденным, но и культурно значимым речевым актом (не случайно существуют специальные предметы клятвы, подчёркивающие культурную значимость этого акта: тфилин[5], перстень, крест). Очень важно тут, что нательный крест носят именно на сердце, поэтому, совершая жест **бить себя в грудь**, жестикулирующий указывает на место, где располагается крест. Ср. также другие русские жесты, связанные с грудью и крестом: **перекреститься** и **скрестить руки на груди**.

Другой смысловой компонент, входящий в толкование жеста – ‘жестикулирующий пытается убедить в чём-то адресата’, – частично реализуется в телесном компоненте «взгляд в глаза адресату». ‘Убеждение’ связано с ‘пониманием’, а по глазам можно прочесть, понимает человек сказанное ему или нет. Взгляд в глаза в русской культуре сопровождает многие семиотические акты, такие как клятва, убеждение, мольба, выражение благодарности.

Интенсивность желания убедить адресата передаётся в этом жесте также и физическим компонентом ‘многократность <исполнения жеста>’. Данный компонент характеризует многие жесты, связанные с интенсивностью, ср. **перебирать пальцами**, **топать ногами, бить поклоны, раскланиваться**. Выражение интенсивности и многократность жестикуляции связаны с высказываниями, которые содержат повторы, например, *Да-да-да!*; *Ни-ни-ни!*; *Приду, приду, приду!*. Кроме того, эти смыслы и передающие их невербальные единицы согласуются со смыслами отдельных естественно-языковых высказываний, обычно со смыслами тех из них, которые сопровождаются силовыми жестовыми ударениями. Это такие высказывания, как *Ну сколько же раз тебе это можно говорить!*; *Раз, два, три!*, и под., а также тексты считалок.

(3) Разные формы тела и его частей могут тоже служить средством поверхностного выражения смысловых компонентов в составе жеста. Проиллюстрируем это утверждение двумя примерами.

(а) Возьмём жест **«Шлагбаум»**. Его физическая реализация такова: одна или обе выпрямленные руки стоящего человека вытянуты горизонтально в сторону. Такое положение рук иконически соответствует

[4] Противопоставление жестов, направленных на отправителя сообщения, и жестов, направленных на адресата сообщения, является существенным для устного диалога.

[5] Тфилин – это коробочка, сделанная из цельного куска телячьей кожи, куда иудеи заворачивают полоски пергамента – отрывки из Торы, представляющие собой заповеди, которые верующие евреи по наступлению совершеннолетия должны исполнять в течение дня. Тфилин укрепляют на бицепсе руки, точнее, на выпуклости мускула между локтем и плечом, и этот священный предмет канонически обращён к сердцу. Тфилин как знак завета с Богом может служить предметом клятвы.

066-Pereverzeva-Kreidlin:_Layout 1 18.05.2008 16:04 Page 430

положению закрытого шлагбаума (отсюда и номинация жеста). Форма руки здесь, таким образом, передаёт смысл ‘проход/проезд закрыт’.

(б) Форма «согнутая спина» реализует смыслы ‘уважение/почтение’, ‘унижение’, в разных видах русских **поклонов**, в мусульманском ритуальном поклоне **сужда**, в жесте **преклонять колени** и некоторых других.

(в) Смысл ‘удивление’ передаётся изменением формы и размера глаз. Жест с такой физической реализацией имеет по меньшей мере два названия. Одно из них – *сделать круглые глаза* – акцентирует изменение формы, но не размера глаз; в другом – *широко раскрыть глаза* – отображается именно изменение размера глаз, но не их формы.

Заключение

В настоящей работе мы рассмотрели некоторые проблемы сочетаемости, а именно, сочетаемости отдельных смысловых элементов в толкованиях невербальных единиц с теми или иными телесными компонентами, входящими в их физическую реализацию. Было показано, что правила перехода от смысла к тексту для единиц языка тела должны учитывать возможность реализации отдельных смысловых компонентов отдельными телесными компонентами, в частности, признаками тела и его частей. Такие правила позволяют уточнить и дополнить правила внутреннего синтаксиса и внешнего синтаксиса, в частности, правила согласования знаковых единиц вербального и невербального кодов.

Список литературы

1. Крейдлин 2002 – Крейдлин Г. Е. Невербальная семиотика: Язык тела и естественный язык // М.: Новое литературное обозрение, 2002.

2. Крейдлин 2006 – Крейдлин Г. Е. Механизмы взаимодействия невербальных и вербальных единиц в диалоге I: Жестовые ударения // Труды международной конференции «Диалог 2006»: компьютерная лингвистика и интеллектуальные технологии». М., 2006. С.290-296.

3. Крейдлин 2007 – Крейдлин Г. Е. Механизмы взаимодействия невербальных и вербальных единиц в диалоге: II А. Дейктические жесты и их типы // Труды международной конференции «Диалог 2007«: компьютерная лингвистика и интеллектуальные технологии». М., 2007. С.300-327.

4. Крейдлин 2008 – Крейдлин Г. Е. Механизмы взаимодействия невербальных и вербальных единиц в диалоге: II Б. Дейктические жесты и речевые акты // М. 2008 (в этом томе).

5. Крейдлин, Переверзева 2008 – Крейдлин Г. Е., Переверзева С. И. Признак «ориентация части тела» в семиотической картине мира // Красильникова Е.В. (отв. ред.). Сборник научных работ в честь А. Б. Пеньковского. М., 2008 (в печати).

067-Kedrova-Potemkin:_Layout 1 18.05.2008 16:04 Page 431

ВЫРАВНИВАНИЕ НЕРАЗМЕЧЕННОГО КОРПУСА ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ

ALIGNMENT OF UN-ANNOTATED PARALLEL CORPORA

Потемкин С.Б. *(potemkin@philol.msu.ru),* ***Кедрова Г.Е.*** *(kedr@philol.msu.ru)*
Московский государственный университет им. М.В. Ломоносова

Предлагается два связанных между собой алгоритма выравнивания параллельных текстов на уровне предложений и на пословно-пооборотном уровне. Метод основан на сопоставлении редко встречающихся слов в обоих текстах с последующим пременением динамического программирования.

Введение

Выравнивание параллельных текстов, то есть автоматическое сопоставление предложений или слов в одном тексте их эквивалентам в переводе, является очень важным этапом предварительной обработки для многих приложений, включая, помимо прочего, машинный перевод (Brown и др., 1993), информационный поиск по текстам на различных языках (Hiemstra, 1996), составление словарей (Smadja и др., 1996) и получение данных для обработки естественного языка (Kuhn, 2004), создание корпуса параллельных текстов (Гарабик, Захаров, 2006).

Наиболее широко применяемые статистические методы, которые не требуют развитой словарной базы и могут использоваться для редких языков, часто дают ошибочные результаты выравнивания, требуя в последующем дорогостоящей ручной проверки и исправления. Использование двуязычных словарей для выравнивания текстов менее распространено, и применялось в основном для специализированных текстов, в то время как для художественных текстов, в которых гораздо чаще встречаются неоднозначные соответствия между предложениями, примеры применения метода приводятся редко.

В отличие от известных методик, в качестве элементов матрицы смежности рассматриваются не отдельные сопоставленные слова, а интервалы от одного пробела в предложении до следующего. Это дает возможность сопоставлять словарные словосочетания из одного предложения – слову или словосочетанию из другого.

1. Предшествующие работы

Выравнивание по предложениям

Методы выравнивания по предложениям относятся, грубо говоря, к трем категориям:

- На основании длины (Gale и Church, 1993), в предположении, что длина предложения в оригинале и в переводе примерно совпадают.
- На основании двуязычной лексической информации, например полученной из корпуса (Kay и Ruoscheisen, 1993; Fung и Church, 1994; Fung и McKeown, 1994).
- Алгоритмы с привлечением опорных меток выравнивают предложения на основании информации, содержащейся в размеченном корпусе или по орфографическому сходству (Simard и др., 1992).

Методы на основании длины очень чувствительны к пропускам в том смысле, что отдельный пропуск может приводить к неправильному последующему выравниванию от точки пропуска до конца корпуса.

Для вычисления сходства двух структурных единиц текстов вводится некоторая мера близости, например количество переводных эквивалентов, имеющихся в словаре. Полученный вес нормализуется на длину текста, чтобы величины для разных единиц текста были сопоставимы. (А.Ф. Гельбух и др., 2006).

Для оптимального решения задачи выравнивания применяется метод динамического программирования. Однако вычисление всей матрицы сопоставления для достаточно больших текстов не представляется возможным вследствие больших временных затрат. Поэтому вместо решения задачи на всей матрице предлагается решать её на некоторой окрестности диагонали (Липатов, Мальцев, 2006).

067-Kedrova-Potemkin:_Layout 1 18.05.2008 16:04 Page 432

Выравнивание по словам

Выравнивание на уровне ниже уровня предложений обычно выполняется с использованием статистических моделей для машинного перевода (Brown и др., 1993; Hiemstra, 1996; Vogel и др., 1999) где любое слово целевого языка считается возможным переводом для каждого слова исходного языка. Для каждой пары слов исходного и целевого текстов выписывается число сегментов, которые (a) содержат оба слова, (b) содержат слово исходного языка не содержат слово целевого языка (c) слово целевого языка, но не слово исходного языка и (d) ни то ни другое слово (Ribeiro и др., 2000).

Найденные таким образом наиболее вероятные пары принимаются в качестве переводных эквивалентов. Такой подход имеет ряд недостатков, связанных с большим количеством редких слов, обычным для любого корпуса, различиями в порядке слов в языках, между которыми производится выравнивание, и наличие словосочетаний. Аналогичным образом построен алгоритм (Липатов, Мальцев, 2006), но с использованием имеющегося англо-русского словаря. Наилучшая пара эквивалентов отыскивается в параллельных предложениях локально, без учета оставшейся части предложения. При этом возможно сопоставление одному слову русского языка нескольких эквивалентов английского, и наоборот.

Другая проблема касается различий в порядке слов между исходным и целевым языком. В предположении локальности инверсии в порядке, который может быть оформлен в пределах некоторого окна, возможна обработка этого явления (Потемкин, Кедрова, 2007).

Подводя итог, отметим, что выравнивание, как на уровне предложений, так и на более низком уровне (слова, словосочетания), нуждается в совершенствовании: существующие модели не позволяют выравнивать тексты с пропущенными или несовпадающими предложениями, редкие слова и словосочетания в пределах предложений. Синтаксические различия между исходными и целевыми языками также представляют затруднения для большинства стратегий выравнивания.

2. Предлагаемый метод выравнивания по предложениям

Основной проблемой при автоматическом выравнивании текста на уровне предложений является появление ложных пар предложений, полученных при работе алгоритма, но не являющихся переводными эквивалентами. При разработке предлагаемого метода мы старались свести к минимуму такие явления.

Алгоритм выравнивания разработан в предположении, что (a) порядок предложений в русском и английском текстах совпадает (b) в параллельных текстах нет значительных (более 500 слов) пропусков (c) длина текстов не слишком большая – рассказ или глава романа, около 64. Последние ограничения непринципиальны и связаны в основном с временем работы алгоритма. Метод основан на использовании обширного англо-русского словаря (Кедрова, Потемкин, 2005), по которому выполняется поиск переводных эквивалентов из анализируемых текстов. В отличие от статистических методов и методов, основанных на мере близости, предлагается рассматривать только низкочастотные слова, а именно слова, встречающиеся только 1 раз в каждом тексте (*hapax legomena*). Вначале для каждого такого слова (русского) текста находим переводной эквивалент в (английском) тексте, который также имеет частотность 1. Если для этого русского слова нашлось несколько эквивалентов, все они исключаются из рассмотрения. Если, далее, найденные эквиваленты связывают предложения с нарушением их порядка в тексте, они также исключаются. В результате такой ограничительной стратегии получаем набор уникальных пар эквивалентов в двух текстах. Такие пары образуют первичную структуру опорных точек или якорей, связывающих те предложения текстов, к которым они относятся. На этом этапе мы не говорим об эквивалентности найденных пар предложений, можно только утверждать, что такие пары предложений имеют непустое пересечение. Затем исходные тексты разбиваются на отрезки, ограниченные найденными парами предложений. Эти отрезки рассматриваются как новые параллельные тексты и процедура расстановки опорных точек повторяется. Итерации продолжаются, пока появляются новые якоря. На практике число итераций в рассмотренных текстах не превышает 6.

После определения опорных точек производится поиск критического пути методом динамического программирования. Для этого рассматриваются отрезки русского и английского текста между опорными точками. Для каждого слова отрезка русского текста отыскивается словарный эквивалент в соответствующем отрезке английского текста. Число таких эквивалентов подсчитывается для каждой пары предложений. Полученные таким образом меры сходства нормируются на единицу по длине предложений и записываются в матрицу смежности. К элементам матрицы, соответствующим опорным точкам к значению меры сходства прибавлено большое число (10000), чтобы критический путь заведомо прошел через эти точки. Далее поиск критического пути выполнялся стандартными методами динамического программирования («поиск Витерби»).

067-Kedrova-Potemkin:_Layout 1 18.05.2008 16:04 Page 433

3. Эксперименты по выравниванию на уровне предложений

Для оценки работы этой части алгоритма было произведено выравнивание нескольких текстов произведений русской классики (Н.В. Гоголь, Ф.М. Достоевский, А. П. Чехов) На рис. иллюстрируются результаты эксперимента, проведенного для рассказа А.П. Чехова «Анна на шее» и его перевода на английский язык. Тексты не подвергались никакой предварительной обработке или разметке. В качестве границы предложения приняты точка, восклицательный знак, вопросительный знак, многоточие.

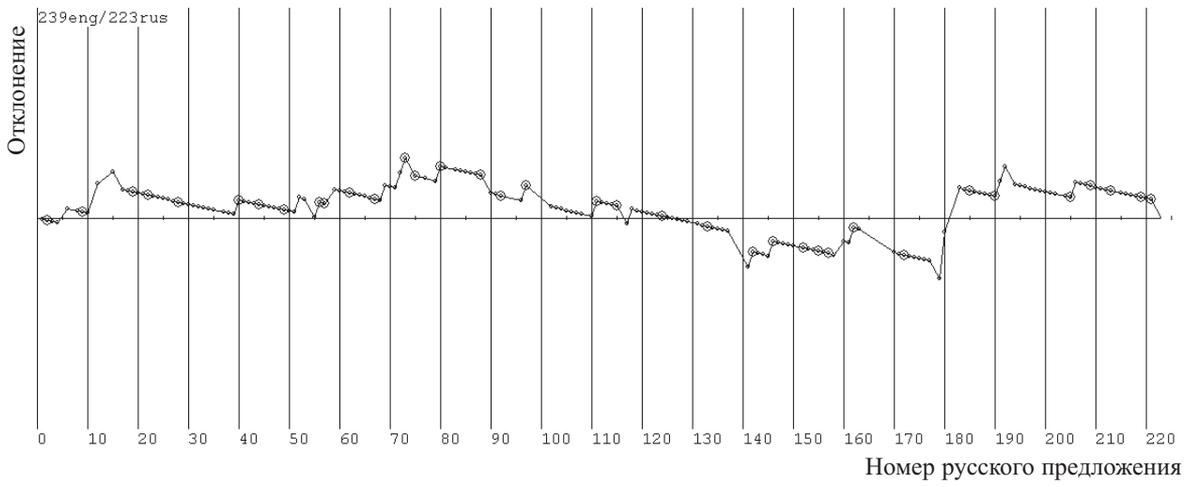


Рис. 1 Диаграмма выравнивания (Гоголь, «Шинель)
Показаны точки критического пути. Опорные точки выделены кружками.

Горизонтальная линия представляет среднее = отношение числа английских предложений к числу русских предложений. По оси Y показано отклонение номера английского предложения от среднего.

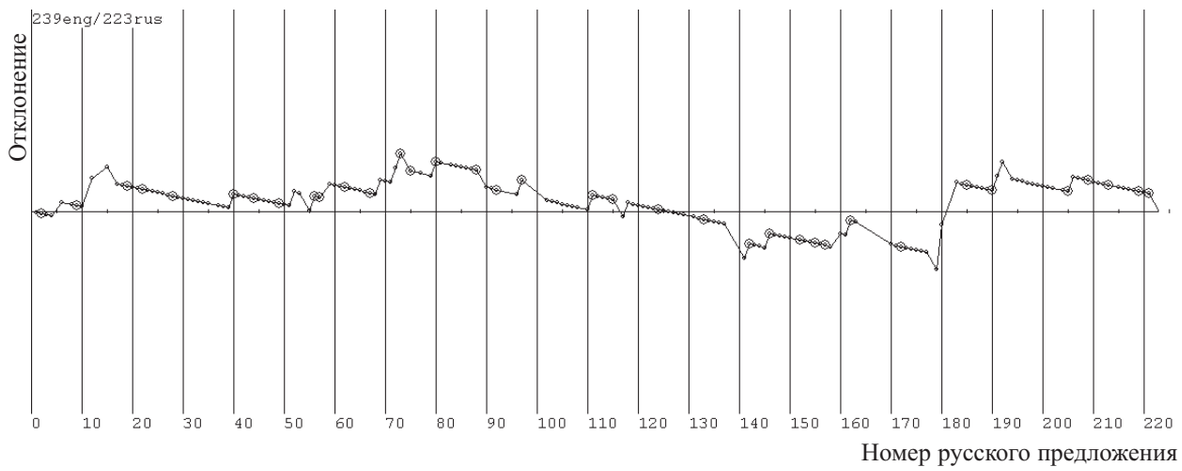


Рис. 2 Диаграмма выравнивания (Чехов, «Анна на шее»)

При выделении предложений учитываются некоторые распространенные сокращения, содержащие точку (Mrs. Mr. Ms. Prof. Dr. Gen. Rep. Sen. St. etc. i.e. e.g. et al.; т.д. т.к. пр. Св. и другие). Эти сокращения распознаются, и точка в этом случае не считается концом предложения. Не учитываются символы перевода строки, заглавные буквы, двоеточия, кавычки. Отказ от учета перевода строки связан с тем, что многие тексты в электронном виде получены в результате сканирования и перевод строки в этом случае может не совпадать с авторским делением на абзацы. Кроме того, в переводе деление на абзацы, также как выделение прямой речи и расстановка других знаков препинания, как правило, отличается от оригинала.

В результате работы алгоритма получено 182 пары предложений (78% текстов). Из них 165 предложений (90.5%) являются полным и точным переводом, 16 предложений (9%) являются частью перевода оригинала (или

067-Kedrova-Potemkin:_Layout 1 18.05.2008 16:04 Page 434

наоборот) и 1 предложение (0.5%) сопоставлено переводу ошибочно. Аналогичные соотношения сохраняются для других текстов («Шинель» Гоголя, главы из романа «Преступление и наказание» Достоевского).

Типичный пример сопоставления части предложения целому:

*«**Статский советник...** принят у его сиятельства...« или: «Со средствами ...*

=

visits at His Excellency's « ; or , «A man of means…

(несопоставленная часть выделена жирным шрифтом)

Расхождение вызвано различной расстановкой знаков препинания в русском и английском тексте, а также элиминацией части текста при переводе.

Обработка таких несовпадений выполняется второй частью алгоритма.

Чаще всего бывает, что 2 последовательных предложения одного текста (i, i+1) сопоставлены 2 предложениям другого текста, не являющимся последовательными (j, j+2). (Рис. 3)

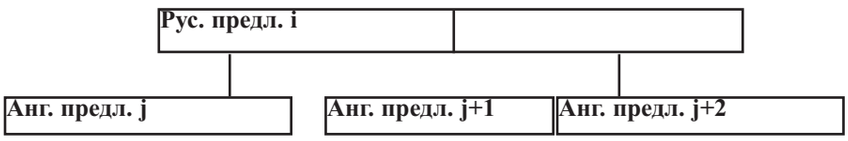


Рис. 3 Объединение 2 предложений при выравнивании

Тогда предполагается, что предложение j+1 является переводом либо части предложения i, либо i+1. Для определения, к какому именно предложению относится j+1 применяется процедура, аналогичная первоначальной разметке, то есть отыскиваются уникальные слова в предложении i, i+1, j+1, их переводы и определяется соответствие, после чего предложение j+1 сливается либо с j, либо с j+2.

Результаты, полученные на основе нашего алгоритма сравнивались с аналогичными (ABBYY, 2007) на тексте «Шинель» Гоголя (433 пары предложений). Доля предложений, правильно сопоставленных алгоритмом ABBYY составила 71%.

4. Фрагментация параллельных предложений

В качестве координатных отсчетов пространства билингвы будем принимать не слова как таковые, а разделители (пробелы) между соседними словами (Потемкин, Кедрова, 2005). При таком подходе отображение слова исходного предложения на слово целевого предложения представляет собой отрезок с координатами начала и конца слова SS по x и начала и конец слова TS по y. Теперь возможно ставить в соответствие не только однословные эквиваленты, но также эквиваленты типа словосочетаний. На рис. 1 показано отображение SS на TS, выполненное с учетом встретившегося в словаре словосочетания (*вдруг == all at once*) Коллизия возникает, когда некоторые отображающие отрезки перекрываются по горизонтали или по вертикали (т.е. отображение не однозначное). Например, слово исходного текста к отображается на слова целевого текста *at, to, with*. Чаще всего в коллизии участвуют служебные слова, а также знаки препинания.

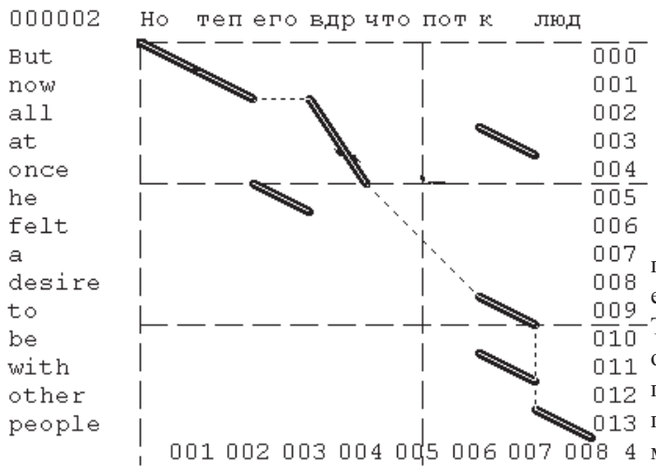


Рис. 4 Отображение SS «Но теперь его вдруг что-то потянуло к людям» на TS «But now all at once he felt a desire to be with other people» (Ф.М.Достоевский) Пунктиром показан критический путь.

Построив пословное отображение, можно переходить непосредственно к фрагментации, то есть к отображению интервалов SS на интервалы TS, которые лежат между уже построенными отображающими отрезками. Слова двух предложений можно сопоставлять в разной последовательности. Одно и то же предложение может быть переведено как с прямым, так и с обратным порядком слов и оба перевода будут

067-Kedrova-Potemkin:_Layout 1 18.05.2008 16:04 Page 435

правильными. Более общий случай – когда одни группы слов переведены в прямом направлении, другие – в инверсном и сами эти группы сопоставляются хаотически.

Данную задачу можно свести к задаче нахождения максимального паросочетания на двудольном графе. Такая задача решается методом нахождения максимального потока на сети или методом, разработанным специально для сетей с единственным истоком и единственным стоком (Диниц, 1970). К сожалению, эксперименты по применению этого метода в настоящее время не привели к положительным результатам по фрагментации предложений с несовпадающим порядком слов в SS и TS.

В случае, если мы рассматриваем только монотонные отображения (т.е. считаем порядок слов исходного и целевого предложения по большей части совпадающим), задача попадает в класс более простых задач динамического программирования.

Как правило, исходное предложение и его перевод, даже имеющие в целом совпадающий порядок слов, содержат фрагменты с инверсией, напр. *{изредка только; only occasionally}*. Такую частичную инверсию желательно включить в критический путь, но приведенный алгоритм этого не допускает. Предлагается формировать фиктивное отображение для инверсных фрагментов заданной длины (напр., не больше 3 слов), которые включаются в общий набор отображающих отрезков и участвуют в алгоритме поиска критического пути.

Вернемся теперь к предложению рис. 1.

Критический путь разбивает исходную пару предложений на следующие фрагменты:

1. Но теперь == But now
2. теперь его вдруг == now all at once
3. вдруг что-то потянуло к == all at once he felt a desire to be with
4. к людям == with other people

Фрагмент 1 не вызывает возражений. Фрагмент 2 SS содержит местоимение его, которого нет во фрагменте TS. Наоборот, фрагмент 3 TS содержит местоимение he, которое отсутствует во фрагменте SS. Если мы объединим фрагменты 2 и 3, результат будет более осмысленным. При решении вопроса об объединении фрагментов мы придерживаемся двух критериев: a) отношение длин фрагментов SS и TS не должно сильно отличаться от 1 и b) переводы всех слов, содержащихся в SS должны содержаться в TS и наоборот. Фрагмент 4 является конечным или хвостовым, и уже не может быть объединен со следующим и его оценка по любому из критериев a) или b) не производится. Поэтому доверие к хвостовому фрагменту заведомо ниже, чем к начальным.

Эксперименты проводились на корпусе параллельных литературных текстов и текстов юридического содержания, на котором метод дает гораздо лучшие результаты. В развитие данного подхода будет составлен автоматический словарь фрагментов для использования в системе автоматического перевода, основанного на примерах (EBMT) (Потемкин, Кедрова, 2007).

Список литературы

1. Brown P.F., Della Pietra S.A., Della Pietra V.J., Mercer R.L., 1993. The mathematics of machine translation: Parameter estimation. // Computational Linguistics, 19(2):263–311.

2. Dinic E.A. An algorithm for the solution of the max-flow problem with the polynomial estimation. Dokl. Akad. Nauk SSSR 194 (1970), no.4 (in Russian); English transl.: Soviet Math. Dokl. 11 (1970), 1277-1280

3. Fung P., McKeown K., 1994. Aligning noisy parallel corpora across language groups: word pair feature matching by dynamic time warping. // Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA-94), pages 81–88, Columbia, Maryland, USA.

4. Gale W.A., Church K.W., 1993. A program for aligning sentences in bilingual corpora. // Computational Linguistics, 19(1):75–102.

5. Kay M., Roscheisen M., 1993. Texttranslation alignment. // Computational Linguistics, 19(1):121–142.

6. Kuhn J., 2004. Exploiting parallel corpora for monolingual grammar induction – a pilot study. // Workshop proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), pages 54–57, Lisbon, Portugal. LREC Workshop: The Amazing Utility of Parallel and Comparable Corpora.

7. Ribeiro A., Lopes G., Mexia J., 2000 A Self-Learning Method of Parallel Texts Alignment // J.S. White (Ed.): AMTA 2000, LNAI 1934, pp. 30−39, 2000.Springer-Verlag Berlin Heidelberg 2000

8. Schrader B., 2006 ATLAS – a new text alignment architecture // Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 715–722, Sydney, July 2006. Association for Computational Linguistics

067-Kedrova-Potemkin:_Layout 1 18.05.2008 16:04 Page 436

9. Simard M., Foster G.F., Isabelle P., 1992. Using cognates to align sentences in bilingual corpora. // Proceedings of the Fourth International conference on theoretical and methodological issues in Machine translation, pages 67–81, Montreal, Canada.

10. Vogel S., Ney H., Tillmann C., 1999. HMM-based word alignment in statistical translation. // Proceedings of the International Conference on Computational Linguistics, pages 836–841, Copenhagen, Denmark.

11. ABBYY, 2007 Выравниватель предложений текстов на русском и английском языке, http://webaligner.abbyy.com/

12 Гарабик Радован, Захаров Виктор Параллельный русско-словацкий корпус // Труды международной конференции Корпусная лингвистика – 2006. Sankt-Petersburg: St. Petersburg University Press 2006

13. Гельбух А.Ф., Сидоров Г.О., Вера-Феликс А., 2006. Словари в задачах автоматической обработки пар переводных текстов // Труды межд. конференции Диалог-2006, М., стр. 110-114.

14. Кедрова Г.Е., Потемкин С.Б., 2004 Семантическое разделение омонимов с использованием двуязычного словаря и словаря синонимов, // II Международный конгресс исследователей русского языка «Русский язык: исторические судьбы и современность», Доклады, М 2004

15. Крылов С.А., Старостин С.А., Актуальные задачи морфологического анализа и синтеза в интегрированной информационной среде STARLING // Труды международной конференции Диалог'2003, М.2003

16. Липатов А.А., Мальцев А.А., 2006 Методы автоматизации построения и пополнения двуязычных словарей с использованием корпуса параллельных текстов // Труды международной конференции Диалог'2006, М.2003

17. Потемкин С.Б., Кедрова Г.Е., 2005 Автоматическая оценка качества машинного перевода на основе семантической метрики // Труды II Международной научно-практической конференции, посвященной Европейскому Дню языков, Луганск 2005.

18. Потемкин С.Б., Кедрова Г.Е., 2007 Использование корпуса параллельных текстов для пополнения специализированного двуязычного словаря // III Международный конгресс исследователей русского языка «Русский язык: исторические судьбы и современность», Доклады, М 2007

068-Prozorova_E:_Layout 1 18.05.2008 16:04 Page 437

ТРАНСКРИПЦИЯ КАК СРЕДСТВО АНАЛИЗА ПАУЗ В РУССКОМ ЖЕСТОВОМ ДИСКУРСЕ[1]

TRANSCRIPTION AS A TOOL FOR ANALYSIS OF PAUSES IN RUSSIAN SIGN LANGUAGE DISCOURSE.

Прозорова Е.В. *(zhenia-pr@yandex.ru)*
Московский государственный университет им. М.В. Ломоносова

Работа посвящена анализу пауз в русском жестовом дискурсе. Для выявления и описания различных типов пауз используются данные транскрипции жестового дискурса: информация о последовательных фазах выполнения жеста, об изменениях выражения лица и положения тела говорящего.

1. Введение

Данная работа посвящена анализу такого просодического явления в дискурсе русского жестового языка, как паузы. Русский жестовый язык (РЖЯ) – это язык, которым в повседневном бытовом общении пользуются глухие люди на территории России. Информация в РЖЯ кодируется при помощи движения и конфигурации рук. Кроме того, существенная часть информации передается немануальными сигналами: выражением лица, направлением взгляда, изменением формы рта, положением головы и корпуса говорящего.

В жестовых языках глухих используется визуально-пространственный, а не звуковой план выражения. Тем не менее в их структуре, так же как в структуре более привычных лингвистам звучащих языков, выделяется просодический уровень.

Просодические явления звучащих и жестовых языков по своей физической сути резко отличаются друг от друга. Применительно к звучащим языкам под просодией понимается совокупность фонетических признаков, таких как тон, громкость, тембровая окраска, ритм.

Ритм также является одной из важнейших просодических характеристик жестовой речи: он складывается из убыстрения/замедления темпа выполнения жестов, увеличения амплитуды жеста (пути руки при выполнении жеста), длительности жеста и наличия пауз. Выделение или акцентированность отдельных жестов достигается за счет повышения скорости их выполнения (Wilbur 1999) или увеличения длительности жеста (Nespor & Sandler 1999). Функции, аналогичные функциям интонации звучащих языков, в жестовой речи берет на себя немануальный компонент (Nespor & Sandler 1999, Wilbur 2000). Немануальные сигналы осуществляют членение жестовой речи на просодические единицы, указывают на их иллокутивную функцию, выражают различные семантические и прагматические характеристики.

2. Принципы транскрибирования жестовых языков

Жестовые языки глухих являются бесписьменными языками, и любое их исследование предполагает первоначальную фиксацию языкового материала при помощи видео. Для дальнейшей работы с собранным материалом необходимо его затранскрибировать: транскрипция позволяет выделить наиболее существенные компоненты высказывания на жестовом языке, зафиксировать те явления, которые ускользают от наблюдателя даже при многократном просмотре видео.

За счет того, что в формировании жестовой речи одновременно участвуют различные части тела, жестовые языки могут передавать большое количество информации в единицу времени. В работах по жестовым языкам общепринятым стало записывать жестовое высказывание в несколько строк, расположенных друг под другом, что позволяет в удобной форме отражать в транскрипции те явления, которые происходят в жестовой речи одновременно. Стандартная запись содержит строку для транскрибирования действий доминантной руки (более активной, у большинства людей - правой), под нею располагается строка для записи действий недоминантной руки (менее активной, у большинства людей - левой), а над нею – строка, где с помощью диакритики транскрибируются немануальные сигналы.

[1] Исследование поддержано грантом РФФИ № 07-06-00061

068-Prozorova_E:_Layout 1 18.05.2008 16:04 Page 438

Принцип многоуровневой транскрипции лежит в основе программы ELAN (Hellwig 2003), с помощью которой аннотируются и транскрибируются корпуса жестовых текстов в Европе (Nonhebel et al. 2004) и Австралии (Johnston, Shembri 2005). Данная программа используется для транскрибирования текстов и в нашем исследовании.

Транскрипция ELAN состоит из нескольких слоев (tiers). Один слой представляет собой множество записей (аннотаций) значений некоторого компонента жестовой речи. Длина аннотации в транскрипции соответствует длительности аннотируемого явления в видеоисточнике. Это позволяет четко определять, какие аннотации в каждом из слоев соответствуют данному моменту видеозаписи.

Наша транскрипция включает слои для отображения жестов правой руки и жестов левой руки, слой, содержащий свободный перевод с РЖЯ на русский, и слои для фиксации немануальных сигналов, таких как направление взгляда; степень раскрытости глаз; положение (поворот и наклон) головы; движение головы; положение подбородка; положение бровей; форма щёк; форма рта.

Используя многослойную транскрипцию в исследовании пауз в дискурсе РЖЯ, мы можем установить, что в момент паузы происходит на всех уровнях жестового высказывания: какое положение занимают руки, о чем говорит выражение лица, положение тела и направление взгляда говорящего.

3. Определение фаз жестов

На начальных этапах транскрибирования жестового дискурса мы обнаружили, что не все жесты структурно одинаковы. Некоторые жесты в потоке жестовой речи обладают ярко выраженным «началом», когда рука говорящего из положения покоя (будучи зафиксированной у груди говорящего; лежа на коленях, на столе или другой поверхности) или из места локализации предыдущего жеста перемещается в место локализации последующего жеста и только тогда приступает к выполнению жеста. Или наоборот – после выполнения жеста рука может возвращаться в положение покоя, тогда жест обладает явно выраженным «концом». Но есть и жесты, которые будто «накладываются» друг на друга – рука приступает к выполнению следующего жеста, не успевая закончить предыдущий.

Наличие разных со структурной точки зрения типов жестов создает определённый ритмический рисунок жестовой речи. Для его отображения мы ввели в транскрипцию слои для записи фаз жестов (отдельно для правой и левой руки).

При определении фаз жестов мы руководствовались работой (Kita et al. 1998). Согласно этой работе, жест (жестовая фраза в терминологии авторов) состоит из обязательной фазы реализации и факультативных фаз: фазы подготовки, которая предшествует фазе реализации, и фазы угасания, которая за ней следует. Сама фаза реализации может быть представлена фиксацией руки в неподвижном положении и неизменной конфигурации (independent hold); или характеризоваться резким движением (stroke), которое может факультативно предваряться и завершаться дополнительными фиксациями руки (dependent holds).

В ходе фазы подготовки рука принимает нужную конфигурацию и занимает то положение, где будет реализован жест. Чаще оба этих процесса происходят одновременно, но иногда рука сначала перемещается в место выполнения жеста и только потом меняет конфигурацию, в соответствии с чем в фазе подготовки выделяют две подфазы: подготовка локализации и подготовка конфигурации.

Как будет показано далее, появление на конце жеста факультативных фаз, особенно фазы фиксации руки и фазы угасания, воспринимается носителями языка как пауза между данным и последующим жестом.

4. Понятие паузы в жестовом дискурсе

В создании дискурсивной транскрипции большую роль играет выявление пауз и измерение их длительности (DuBois et al. 1992; Кибрик, Подлесская 2003). Паузы существенны для понимания когнитивных процессов, сопровождающих процесс порождения речи.

В устной речи пауза определяется как перерыв в артикуляции и соответствующий ему физический перерыв в речевом (звуковом) сигнале (Кривнова, Чардин 1999). В буквальном виде это определение неприменимо к жестовым языкам, где сигнал обладает не звуковой, а визуально-кинетической природой. Однако жестовая речь, как и устная, имеет хорошо воспринимаемый и последовательно воспроизводимый ритм (Allen et al. 1991), который задается, главным образом, перерывами, нарушениями плавного, однообразного течения жестовой речи. Такие перерывы и определяются как паузы в работах по жестовым языкам (например, Grosjean & Lane 1977)

068-Prozorova_E:_Layout 1 18.05.2008 16:04 Page 439

Вместе с тем, в известных нам исследованиях не рассматривается подробно, что является физическим выражением паузы в жестовом дискурсе. Определение паузы как отрезка между двумя жестами, во время которого руки говорящего расслаблены, не имеют четко выраженной формы и локализации, приведённое в работе Nespor & Sandler (1999), не охватывает разнообразных типов пауз в дискурсе РЖЯ, которые будут рассмотрены далее.

Мы покажем, что паузы в РЖЯ характеризуются не только действиями рук, но и немануальными сигналами, и предложим пробную классификацию пауз.

5. Методы и материал исследования

Исследование проводилось на материале записанных на видео нарративов на русском жестовом языке. Всего было проанализировано 5 текстов общей длительностью 5 минут 6 секунд. Каждый текст представляет собой пересказ сюжета одного из комиксов художника Х. Бидструпа. Носителям языка, участвовавшим в съемках, предъявлялась картинка, которую надо было внимательно рассмотреть, запомнить, а потом, не глядя на неё, пересказать сюжет другому глухому участнику. Всего в съемках участвовало 5 рассказчиков, 4 женщины и 1 мужчина в возрасте от 23-х до 37-ми лет.

Расстановка пауз в записях РЖЯ производилась слышащим исследователем совместно с глухим носителем языка. Нарративы РЖЯ демонстрировались с экрана компьютера при помощи программы ELAN в формате Media Synchronisation Mode. Носителю языка выдавалась распечатка транскрипции записи, представленной в виде последовательности жестов (для передачи жеста использовалось наиболее близкое по смыслу русское слово, записанное заглавными буквами), записанных в столбец один под другим. Задание было сформулировано следующим образом: отметить на бумаге те места, где в речи рассказчика наблюдается остановка, заминка или перерыв (даже очень краткий), используя цифры: 1 – для краткой паузы; 2 – для паузы средней длительности; 3 – для долгой паузы. Сначала запись демонстрировалась целиком, после по фрагментам, и расстановка пауз производилась в ходе совместного обсуждения записи исследователем и носителем языка.

Над каждой записью такая процедура производилась дважды, с участием двух разных носителей языка. В 92% случаев мнения носителей о наличии паузы в определённом моменте жестового дискурса совпали. Далее в работе анализировались только эти случаи - всего 164 паузы в 5 текстах.

6. Расположение пауз в жестовом дискурсе

Мы ожидали, что паузами в РЖЯ, как и в звучащих языках, считаются те моменты, когда «ничего не происходит». В жестовых языках для этого должны одновременно выполняться следующие условия: а) руки находятся в неподвижном положении; б) руки расслаблены, не имеют четкой конфигурации; в) не меняется выражение лица и положение тела. Кроме того, казалось вероятным, что говорящий моргает именно в моменты пауз, чтобы необходимость прервать визуальный контакт с окружающим миром минимальным образом затрагивала процесс коммуникации.

Предположение о том, что говорящий с большей вероятностью моргает в момент паузы не подтвердилось на материале РЖЯ. Из 164 отрезков дискурса РЖЯ, охарактеризованных носителями как паузы, только 64 (39%) сопровождались морганием. При этом в разных случаях момент моргания приходился на разные «этапы» паузы – он мог предварять паузу, появляться в начале паузы, сопровождать момент перехода от одного жеста к другому или возникать в финальной части паузы и следовать за паузой. Моргание не является физиологическим регулятором пауз в жестовом дискурсе, каковым является дыхание в устной речи. По всей видимости, моргание не выступает и сигналом перерыва в жестовом дискурсе, а выполняет иные функции, рассмотрение которых выходит за рамки данной работы.

Также анализ записей РЖЯ показал, что вопреки нашим исходным предположениям моменты пауз чаще характеризуются сменой выражения лица и положения головы и тела говорящего, чем отсутствием изменений в мимике и позе. Мы учитывали значения нескольких немануальных сигналов, отображаемых в нашей транскрипции жестового дискурса, а именно: направление взгляда, степень раскрытости глаз, положение головы (направление наклона и/или поворота), движения головы (кивки, качание) и форму рта. Только 31 (19%) пауза в текстах РЖЯ не сопровождалась сменой значений немануальных сигналов.

Несмотря на ту значительную роль, которую смена немануальных сигналов играет в восприятии пауз в жестовом дискурсе, наличие или отсутствие пауз в первую очередь определяется действиями рук говорящего.

068-Prozorova_E:_Layout 1 18.05.2008 16:04 Page 440

7. Определение пауз через фазы жеста

Полученные данные показали, что в жестовом дискурсе пауза не подразумевает полного перерыва в артикуляции. В исследуемых нами записях РЖЯ практически не встретилось таких пауз, когда бы обе руки говорящего находились в положении покоя (лежали на коленях, на столе, занимали неподвижное положение перед грудью говорящего) и одновременно имели бы расслабленную, нечёткую конфигурацию кисти. Такое положение рук характерно, главным образом, для абсолютного начала и абсолютного конца нарратива.

Напротив, в моменты пауз внутри нарратива в большинстве случаев соблюдалось только одно из указанных выше условий: руки либо 1) занимали относительно неподвижное положение, но сохраняли чёткую конфигурацию, либо 2) имели расслабленную конфигурацию, но продолжали движение.

Все случаи типа 1) с точки зрения разбиения жестов на последовательные фазы, представляют собой факультативную фазу фиксации, которая следует за фазой реализации. Напомним, что фиксация руки в неподвижном положении может быть и фазой реализации жеста. Однако в таких случаях носители языка не воспринимали остановку руки как паузу. Далее паузы, во время которых руки говорящего неподвижны, но сохраняют чёткую конфигурацию, мы будем называть паузами типа Ф (Рис.1).

А.  Б. 

Рис.1. Пауза типа Ф, длительность: 148 мсек.
А: начало паузы. Б: конец паузы

Случаи типа 2) приходятся на фазу угасания жеста. Для этой фазы характерно расслабление кисти руки и одновременное перемещение руки из места выполнения жеста в положение покоя. Такие паузы мы будем называть далее паузами типа У (Рис.2)

Количество пауз типа Ф в анализируемых текстах составило 86. Количество пауз типа У – 57. Число пауз, состоящих из двух последовательных фаз фиксации и угасания (паузы Ф+У) составило 7. Как паузы типа Ф, так и паузы типа У могут сопровождаться различными изменениями немануальных сигналов или отсутствием таких изменений.

На начальном этапе исследования нам не удалось выявить зависимость между типом паузы и какими-либо формальными характеристиками контекста, в котором встречаются такие паузы.

Для того, чтобы описать функциональные различия между двумя типами пауз, требуется более детальный анализ, выходящий за рамки данной работы. На данном этапе мы можем только сформулировать предварительную гипотезу. Мы считаем, что появление в жестовом дискурсе пауз типа У может быть обусловлено особыми требованиями жестовых языков к формальному аспекту производства жестов. Например, возвращение руки из точки, в которой рука закончила фазу реализации жеста, обратно к груди говорящего (в центр жестовой области), и только потом перемещение в точку, где рука должна начать выполнение

068-Prozorova_E:_Layout 1 18.05.2008 16:04 Page 441

Рис.2. Пауза типа У, длительность 498 мсек.
А-Б: Рука покидает место реализации жеста. В: Рука занимает положение покоя у груди

последующего жеста, необходимо для того, чтобы жесты не «сливались» в ходе порождения жестовой речи, а имели четкие границы, единый ритмический контур. Таким образом, паузы типа У в жестовом дискурсе наиболее близки к понятию паузы в звучащем дискурсе: они представляют собой перерыв в «артикуляции» жестовой речи, когда ни форма руки, ни её движение, не несут семантической нагрузки.

Напротив, появление в жестовом дискурсе пауз типа Ф, скорее всего, обусловлено не формальными, а когнитивными факторами. Конфигурация руки, сохраняясь после фазы реализации жеста, продолжает нести определённую семантическую нагрузку. Таким образом, несмотря на перерыв в подаче информации, идея, переданная жестом, остаётся активированной для коммуникантов.

8. Заполненные паузы

Некоторые отрезки дискурса РЖЯ, которые носители языка определили как паузы, представляют собой не отдельную фазу жеста, а жест целиком. В рассмотренных нами текстах встретилось два вида таких жестов (выполняются одной рукой):

Рис.3. Пример заполненной паузы

1) рука не перемещается; пальцы выпрямлены, направлены под углом к ладони и совершают «перебирающее» движение (Рис.3); 2) рука не перемещается; указательный палец выпрямлен и направлен вверх.

Всего в текстах встретилось 10 примеров жестов типа 1) и 4 примера жестов типа 2).

Эти жесты не имеют семантического наполнения. Но их появление в жестовой речи свидетельствует о том, что говорящий испытывает определённые затруднения и прикладывает дополнительные усилия, чтобы воспроизвести следующий фрагмент дискурса.

Носители языка воспринимают эти типы пауз как моменты, в которые говорящий «ищет нужное слово» или «пытается что-то вспомнить». Таким образом, функционально эти жесты представляют собой паузы хезитации. С формальной точки зрения их можно сопоставить с заполненными паузами, встречающимися в устном дискурсе – периодами времени, «когда вербализация не происходит, но говорящий производит вокальный сегмент типа редуцированного гласного «шва» (записывается с помощью буквы э) или сонорный носовой сегмент (транскрибируется с помощью буквы м)» (Кибрик, Подлесская 2003:8).

068-Prozorova_E:_Layout 1 18.05.2008 16:04 Page 442

9. Заключение

В нашем исследовании мы сосредоточились на определении понятия паузы для жестового дискурса и постарались описать отрезки жестового дискурса, которые воспринимаются носителями языка как паузы, в терминах фаз жестов и значений немануальных сигналов, которые отображаются в используемой нами системе транскрипции жестового дискурса.

В дискурсе РЖЯ как паузы могут восприниматься отдельные фазы жестов, а именно фаза фиксации, следующая за фазой реализации жеста, а также фаза угасания. Во время фазы фиксации (паузы типа Ф) рука сохраняет конфигурацию жеста, которая обладает определённой семантикой. Во время фазы угасания (паузы типа У) рука, теряя четкую конфигурацию, продолжает движение, хотя, как представляется, в большинстве случаев это движение обусловлено формальными требованиями к выполнению жестов.

Что касается немануального компонента жестовой речи, то момент паузы чаще приходится на резкую смену значений одного или нескольких немануальных сигналов.

В дискурсе РЖЯ также встречаются паузы, которые представляют собой отдельные жесты, указывающие на затруднения, которые испытывает говорящий при порождении жестовой речи.

Типология пауз дискурса РЖЯ, безусловно, требует уточнения и дальнейшей разработки. Однако уже начальные результаты показывают, что механизмы формирования и восприятия пауз в жестовом и устном дискурсе существенно различаются. Если в устном дискурсе пауза – этот перерыв в артикуляции и физическом сигнале, то в жестовом дискурсе как пауза воспринимается контраст между двумя последовательными отрезками жестовой речи – контраст между движением и отсутствием движения; контраст между четкой и расслабленной конфигурацией руки; смена выражений лица, направления взгляда, положения тела.

Вместе с тем, эти различные явления устного и жестового дискурса имеют схожие функции, что и позволяет использовать по отношению к ним один и тот же термин – пауза.

Список литературы

1. Кибрик А.А., Подлесская В.И. К созданию корпусов устной русской речи: принципы транскрибирования // Научно-техническая информация (серия 2). М.: 2003. № 6. С. 5-11.

2. Кривнова О.Ф., Чардин И.С. Паузирование при автоматическом синтезе речи // Теория и практика речевых исследований (АРСО-99). Материалы конференции. 1999.

3. Allen, G., Wilbur, R. & Schick, B. Aspects of rhythm in American Sign Language // Sign Language Studies. 1991. № 72, P. 297-320.

4. Du Bois J.W., Schuetze-Coburn S., Cumming S., Paolino D. Discourse transcription. Santa Barbara papers in linguistics, 4. Santa Barbara: UCSB. 1992.

5. Chafe, W. Some Reasons for Hesitating // Dechert H.W. and Raupach M. (eds.) Temporal Variables in Speech. The Hague: Mouton, P.169-180. Reprinted in Tannen, D. and Saville-Troike, M. (eds.) Perspectives on Silence. Norwood, NJ: Ablex, 1985. P. 77-89.

6. Hellwig, B. EUDICO Linguistic Annotator (ELAN) version 1.4: manual. Nijmegen, 2003.

7. Grosjean, F. & Lane, H. Pauses and syntax in American Sign Language // Cognition, 1977. № 5, P. 101-117.

8. Johnston T. & Schembri, A. The use of ELAN annotation software in the Auslan Archive/Corpus Project // Presentation at the Ethnographic Eresearch Annotation Conference, University of Melbourne, February 15-16, 2005.

9. Kita, S., van Gijn, I., & van der Hulst, H. Movement Phases in signs and co-speech gestures, and their transcription by human coders // Wachsmuth, I. & Fröhlich, M. (eds.) Gesture and sign language in human-computer interaction, International Gesture Workshop Bielefeld, Germany, September 17-19, 1997. Proceedings. Lecture Notes in Artificial Intelligence Berlin: Springer-Verlag, 1998. Vol. 1317, P. 23-35.

10. Neidle, C., Kegl J., MacLaughlin D., Bahan B., & Lee, R. G. The Syntax of American Sign Language. Cambridge MA: MIT Press, 2000.

11. Nespor, M. & Sandler, W. Prosody in Israeli Sign Language // Language and Speech, 1999. № 42 (2-3), P. 143-176.

12. Nonhebel, A., Crasborn, O. & van der Kooij, E. Sign language transcription conventions for the ECHO project. ECHO project, University of Nijmegen, 2004.

13. Wilbur, R.B. Phonological and prosodic layering of nonmanuals in American Sign Language // Lane, H. & Emmorey, K. (Eds.) The signs of language revisited: Festschrift for Ursula Bellugi and Edward Klima. Hillsdale, NJ: Lawrence Erlbaum, 2000. P. 213-241.

069-Protasov:_Layout 1 18.05.2008 16:05 Page 443

ВЫВОД И ОЦЕНКА ПАРАМЕТРОВ ДАЛЬНОДЕЙСТВУЮЩЕЙ ТРИГРАММНОЙ МОДЕЛИ ЯЗЫКА

INFERENCE AND ESTIMATION OF A LONG-RANGE TRIGRAM MODEL

Протасов С.В. *(svp@tj.ru)*
Московский физико-технический институт (Государственный университет)

В докладе описывается простая вероятностная грамматика связей (Link Grammar), известная также, как “Модель дальнодействующих триграмм” (Long-range Trigram Model). Эта вероятностная модель языка расширяет триграммные модели, предсказывая слова не только по двум непосредственно предшествующим словам в предложении, но и потенциально по любой паре стоящих рядом слов, которые лежат внутри этого же предложения. Таким образом, триграммная модель может пропускать менее информативные слова для более точного пргноза. Лежащая в основе “грамматика” есть не более, чем множество пар слов, которые могут быть связаны вместе через несколько разделяющих слов; это множество слов получается автоматически из корпуса текста, используемого для “обучения модели” грамматики. В докладе представлены результаты экспериментов, совершенные на корпусе предложений русского языка.

1. Вступление

В данной работе мы исследуем модель языка, которая может использоваться для практических задач, где требуется вероятностная оценка корректности предложений. В работе [Protasov 06] автором исследовалась более сложная модель и её реализация не позволила провести обучение (тренировку) на большом корпусе. В частности, использовался корпус около 3 тыс предложений со словарём примерно 300 слов. Конечно же в реальных задачах нам потребуются модели, которые позволяют обрабатывать корпуса размером на несколько порядков больше. Далее мы будем обсуждать одну из таких моделей, которая, по сути, является упрощением контекстно-свободной модели языка из работы [Protasov 06].

Наиболее широко используемой статистической моделью языка в настоящий момент является так называемая *триграммная модель*. В этой простой модели слово предсказывается на основе только лишь двух слов, непосредственно стоящих перед ним. Простота *триграммной модели* одновременно является и ее наибольшим преимуществом, и недостатком. Преимущество модели заключается в том, что для оценки параметров модели языка существует достаточно простой и быстро работающий алгоритм, который может обработать сотни миллионов слов текста. Реализация модели будет содержать внутри всего лишь поиск по большой таблице, что достаточно просто в практическом плане. Все новые статистические модели практически всегда оцениваются по отношению к триграммной модели. На сегодняшний день многие успешные системы распознавания речи в той или иной форме используют именно n-граммную модель (где n=2,3) [Jelinek, 97]. Несмотря на свои успехи триграммная модель ничего не знает о богатых синтаксических и семантических связях, которые содержат естественные языки, позволяя им быть легко распознаваемыми и понимаемыми людьми. Во многих реальных предложениях зависимые слова находятся на довольно большом расстоянии в 5-7 слов и триграммная модель никак не может учесть эти связи. Использование n-граммных моделей с n=5,6,7 требует гиганских ресурсов и сталкивается с проблемой “редких данных”.

Вероятностная грамматика связей была предложена как подход, который сохраняет достоинства и вычислительные преимущества триграммной модели, и в то же время включает дальнодействующие зависимости и более сложную информацию в статистическую модель [Lafferty et al. 92]. В этом докладе будет представлена реализация очень простого варианта *вероятностной грамматики связей,* которая (реализация) применима для любого естественного языка, включая русский. Грамматика расширяет *триграммные модели* через разрешение связей между словами, предшествующими не только в пределах двух предыдущих слов, но и потенциально находящимися на большем расстоянии от предсказываемого слова в пределах предложения. Таким образом *дальнодействующая триграммная модель* может пропускать малоинформативные слова и улучшать предсказуемость в модели. Лежащая в основе грамматика представляет собой множество пар слов, которые могут быть соединены друг с другом через несколько промежуточных слов. Впервые *дальнодействующая триграммная модель* была предложена в работе [Pietra et al. 94], где она исследовалась на англоязычном

069-Protasov:_Layout 1 18.05.2008 16:05 Page 444

материале, но к сожалению результаты исследований ученых из IBM не были подтверждены независимо, не говоря уже о доступности каких-либо программных реализаций модели, а публикации по исследованию на корпусах других языков отсутствуют до сих пор.

Далее во втором разделе будет кратко описано введение в *дальнодействующую триграммную модель* и показано, как она может быть представлена в виде *вероятностной грамматики связи*. Грамматика парных слов автоматически выводится из корпуса обучающего текста. Хотя взаимная информация слов также может использоваться для эвристического вывода парных слов, сам по себе этот подход не приносит адекватных результатов. В третьем разделе будет описан алгоритм, адаптирующий критерий *взаимной информации* для наших целей. В последнем разделе представлены результаты экспериментов, совершенных на русскоязычном материале.

2. Дальнодействующая триграммная модель

В качестве примера рассмотрим рисунок 1. На диаграмме представлена *связка* (linkage) предложения “Если у Вас есть ... заработать.”, согласно формализму, впервые введенному в [Sleator and Temperley, 91], важными свойствами связки является непересечение связей, их связность (отсутствие неприсоединенных областей), единственность связей (каждая пара слов соединена только одной связью). Рассматривая вероятностную модель, мы считаем, что каждое слово генерируется из биграммы, заканчивающейся словом, примыкающим к генерируемому слову слева. Таким образом, первая правая скобка сгенерирована на основе биграммы (сайт|(), а первое слово “сайт” сгенерировано из биграммы (есть|свой). Слово “то” сгенерировано из биграммы (⊥,|Если), где ⊥ является специальным словом-границей.

Рис. 1. Дальнодействующие триграммы

Для описания модели более детально, рассмотрим следующее описание стандартной триграммной модели. Модель может быть рассмотрена как простой конечный автомат, генерирующий предложения. Состояния этого автомата проиндексированы парами слов. Добавив слово-границу ⊥ в наш словарь слов, мы зададим начальное состояние конечного автомата как (⊥,⊥). Когда автомат находится в каком-либо состоянии (w_1,w_2), он может перейти в состояние (w_2,w_3), с вероятностью $t(w_3|w_1,w_2)$ и остановится с вероятностью $t(\perp|w_1,w_2)$, таким образом остановив предложение.

Наша расширенная триграммная модель может быть описана похожим образом. Для ссылки на состояния автомата используются пары слов, но состояние $s = (w_1,w_2)$ теперь может быть одним из трех: останов (halt), шаг (step), ветвление (branch) с вероятностями *d(halt|s), d(step|s), d(branch|s)* соответственно. В случае выбора состояния *step* или *branch*, следующее слово *w* генерируется с триграммной вероятностью $t(w|w_1,w_2)$. Но в случае выбора *branch* генерируется дополнительное слово *w′* на основе дальнодействующей триграммы $l(w'|w_1,w_2)$. Например, в процессе генерирования связки из примера выше, состояние с индексом *s* = (то,регистрация) приводит к состоянию step с вероятностью *d(step|s)* и слово “позволит” затем генерируется с вероятностью *t*(позволит|то,регистрация). С другой стороны, состояние s = (⊥,Если) ответвляется с вероятностью *d(branch|s)* и затем из этого состояния генерируется слово “у” и слово “то” с вероятностью *t*(у| ⊥Если) и *l*(то| ⊥Если).

В результате все слова в связках, как на примере выше, имеют ровно одну связь слева и ноль, одну или две связи справа. Если мы пронумеруем слова в предложении *S* от 1 до |*S*|, тогда вполне удобно обозначать через $\lhd i$ индекс слова, которое генерирует слово слева от i-го в предложении. Таким образом, *i* соединено слева с $\lhd i$. Например, на связке из примера выше мы видим, что $\lhd 9 = 8$, $\lhd 8 = 1$, и $\lhd 26 = 18$. Подобная запись позволяет нам записать вероятность предложения как $P(S) = \Sigma_{L(S)} P(S,L)$, где *L(S)* есть набор всех связок *S* и где соединяющая вероятность *P(S,L)* расписывается как

$$P(S,L) = \prod_{i=1}^{|S|} d(d_i|w_i,w_{i-1})\, t(w_i|w_{i-2}w_{i-1})^{\delta(i-1,\lhd i)}\, l(w_i|w_{\lhd i-1}w_{\lhd i})^{1-\delta(i-1,\lhd i)} \quad (1)$$

069-Protasov:_Layout 1 18.05.2008 16:05 Page 445

Здесь $d_i \in halt, step, branch, \delta(i, j)$ равен единице, если $i = j$, и нулю, если не равен. Индекс $\triangleleft i$ должен пониматься по отношению к заданной связке L.

В терминах *грамматики связей* [Sleator and Temperley, 91] переменные *halt, step* и *branch* эквивалентны трем простым *дизъюнктам*, определяющим, как заданное слово соединяется с другими словами. Значение *halt* соответствует дизъюнкту, имеющему один левый коннектор (без метки) и не имеющий правых коннекторов. Значение *step* соответствует дизъюнкту, имеющему единственный левый и единственный правый коннектор. Значение *branch* соответствует дизъюнкту имеющему один левый коннектор и два правых коннектора. В формализме данной грамматики вероятностная модель (1) является простым вариантом более общей вероятностной грамматики связей, представленной в работе [Lafferty et al. 92].

На этом мы закончим сверхкраткое введение в дальнодействующие триграммные модели и за дополнительной информацией рекомендуем обратиться к работе [Pietra et al. 94]. Там же дано описание эффективного алгоритма “обучения” модели (что равносильно выводу грамматики). Целью алгоритма является увеличение суммы (1) по всем предложениям в обучающем корпусе. Алгоритм “обучения” хоть и является разновидностью EM (Expectation-maximization, разновидность алгоритма максимизации правдоподобия) [Baum 72], в действительности довольно сильно отличается от популярного подхода Inside-Outside [Lari and Young, 90], который часто используется для обучения формальных вероятностных моделей [Manning and Shutze, 99].

3. Вывод грамматики

Вероятностная модель (1), описанная в предыдущем разделе, делает свои предсказания на основе как обычных триграммных моделей, так и на основе дальнодействующих триграмм. Мы можем разрешить использовать связи со словами, присоединяющееся слева к любому слову. Это соответствует “грамматике”, которая разрешает дальнодействующие связи между любыми двумя словами. Число возможных *связок* для такой грамматики растет очень быстро с увеличением длины предложения: если предложение состоящие из 10 слов имеет всего-лишь 835 *связок*, то предложение, состоящее из 25 слов уже имеет 3 192 727 797 *связок*. Однако большинство дальнодействующих связей в этих связках скорее всего будут неправильными. Получившаяся вероятностная модель имеет слишком много параметров, которые не могут быть достаточно точно оценены. А для целей качественного обучения нам требуется высокое отношение “число примеров/число параметров”.

L	R	$log(Gain_{LR})$	$d(branch_{LR}\|L)$	$d_{LR}(halt)^{-1}$
(	)	11.05	0.8558	4.4
Если	то	8.81	0.3541	7.1
либо	либо	8.44	0.3398	4.2
"	"	8.33	0.2171	7.9
Ни	ни	7.92	0.4228	2.6
Чем	тем	7.78	0.4414	5.0
столько	сколько	7.76	0.2661	2.6
Чем_больше	тем	7.66	0.9585	4.2
Что_касается	то	7.47	0.7549	3.5
ни	ни	7.43	0.2123	2.8
Чем_больше	тем	7.66	0.9585	4.2
Ни_одна	не	5.92	0.9364	2.8
Чем_дольше	тем	5.83	0.9157	4.2
кроме_тех_случаев	когда	5.14	0.9241	1.2
только_в_том_случае	если	7.21	0.7349	1.1
Никакой	не	5.22	0.7549	3.1
Что_касается	то	7.47	0.7549	3.5
Интересно	?	5.73	0.2437	8.8
Даже_если	все_равно	5.25	0.1187	8.7
Во-первых	во-вторых	6.08	0.1294	8.5
Неужели	?	7.17	0.4026	7.6
Разве	?	6.57	0.2864	7.6
Ах	!	6.54	0.4918	7.4
Если	то	8.81	0.3541	7.1
Почему	?	7.41	0.3296	7.0
Сначала	потом	5.77	0.2168	6.3
Одна	другая	5.04	0.0877	6.1

069-Protasov:_Layout 1 18.05.2008 16:05 Page 446

L	R	$log(Gain_{LR})$	$d(branch_{LR}\|L)$	$d_{LR}(halt)^{-1}$
отличить	от	6.49	0.6775	1.7
не_обращал	внимания	6.12	0.5178	2.1
избавить	от	5.88	0.7158	1.2
споткнулся	упал	5.86	0.3114	2.3
Одним_из	является	5.81	0.4638	4.3
отделить	от	5.68	0.6820	1.8
превратить	в	5.36	0.6560	1.7
Делать	нечего	5.30	0.4220	2.1
прижала	к	5.17	0.6869	1.3
обращать	внимание	4.91	0.2997	2.0
Целью	является	4.84	0.5089	2.1
нашелся	ответить	4.55	0.2091	1.9
Прошло	прежде_чем	4.53	0.1653	3.0
поблагодарить	за	4.48	0.4136	1.5

Таблица 1. Примеры пар слов

Раз неограниченная грамматика непрактична, мы попробуем ограничить грамматику через разрешение только тех дальнодействующих связей, которые приносят наибольшие улучшения в вероятностную модель. В идеале нам нужно автоматически выявлять пары слов, такие как “(” и “)” с дальнодействующими корреляциями, которые могут быть хорошими кандидатами на соединение через дальнодействующую связь. Мы можем поискать такие пары через просмотр слов с высокой взаимной информацией. Но если мы представим, что мы уже включили связи всех ближайших соседей в нашу модель, как в случае модели (1), то у нас не будет точек для связывания слов *L* и *R*, независимо от того, насколько велика их взаимная информация, ведь слово *R* уже хорошо предсказывается непосредственными предшественниками. Вместо этого мы будем искать связи между словами, которые имеют потенциал улучшения модели только по сравнению с обычными короткими связями.

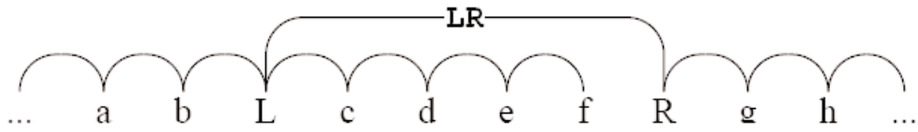


Рис. 2. Модель LR

Для нахождения таких пар используем следующий подход. Пусть V словарь языка. Для каждой пары $(L,R) \in V \times V$ сконструируем модель P_{LR}, которая содержит все связи биграммами с одной дополнительной дальнодействующей связью, идущей от *L* к *R*. На основе анализа корпуса русскоязычных предложений мы определим пользу пары (*L,R*) по сравнению с биграммной моделью.

Мы выбрали модель P_{LR} достаточно простой, чтобы параметры всех $|V|^2$ возможных моделей оценивались параллельно. Затем мы отсортируем модели согласно их правдоподобию $Gain_{LR}$ [Pietra et al. 94], которую каждая модель показывает на обучающем корпусе, и выберем те пары (*L,R*), которые соответствуют самым лучшим моделям. Этот список пар слов и будет составлять нашу новую “грамматику”, описанную в предыдущем разделе.

4. Результаты экспериментов

Этот раздел представляет результаты обучения наших дальнодействующих триграммных моделей на корпусе предложений, собранных через интернет. Наш обучающий корпус состоял из более чем 11 млн. предложений, содержащих примерно 150 млн. слов. Таблица 1 включает примеры пар слов, которые были получены после использования формул из раздела 3. Напомним, что эти пары были получены при первом шаге обучения грамматики связей, которая позволяет дальние связи между одной фиксированной парой слов. Каждая пара проверяется уменьшением энтропии, которая ее односвязная модель достигает по сравнению с биграммной моделью. В таблице это улучшение показано в 3-м столбце. Мы сразу отсекаем все пары, которые не приводят к уменьшению энтропии. В первой секции таблица содержит пары, которые приводят к наибольшему уменьшению энтропии. Четверый столбец таблицы дает значения вероятности $d(branch_{LR}|L)$. Это значение показывает вероятность, с которой *L* генерирует *R* с некоторого расстояния в соответствии с обучаемой моделью. Вторая

069-Protasov:_Layout 1 18.05.2008 16:05 Page 447

секция таблицы включает примеры пар с высоким значением вероятности $d(branch_{LR}|L)$. Пятый столбец таблицы дает значения вероятности $d_{LR}(halt)^{-1}$. Поскольку в обучающих данных число слов между L и R убывает геометрически со средним $d_{LR}(halt)^{-1}$, то большое значение в этом столбце указывает, что L и R находятся в среднем на достаточно большом расстоянии. Третья секция таблицы приводит примеры таких пар. В заключение, четвертая секция таблицы показывает пары, где одно из слов является глаголом, и только некоторые из этих пар с наибольшим уменьшением энтропии показаны в таблице.

Мы довольны результатами, так как полученные списки пар практически не содержат мусора, который в избытке появляется при использовании других методов. Однако из-за нечеткости критерия "что есть мусор", нам очень сложно провести численное сравнение. Более того, нам вообще хотелось бы избежать человеческой оценки качества связей и использовать более формальные оценки.

5. Выводы и планы

Полученные данные позволяют сделать вывод, что модель дальнодействующих триграмм представляет собой еще один инструмент корпусной лингвистики. Этот инструмент, в частности, позволяет автоматически устанавливать факт наличия синтаксической связи между словами, не стоящими рядом. Полученная "грамматика пар слов" может быть использована для инициализации более сложных вероятностных моделей. Исследование пар, отфильтрованных по частям речи, может помочь в изучении "дальних" валентностей глаголов, а также составлению списка глаголов, потенциально имеющих большое число валентностей. Было бы интересно изучить таким способом какой-либо мертвый язык, имеющий достаточно большой корпус текстов. Однако наша долгосрочная цель не словарь парных слов, а более мощная статистическая модель языка. Если в процессе тренировки модели мы получаем качественный словарь, содержащий мало мусора, то это хорошее свидетельство того, что мы движемся в правильном направлении. После того как мы получили список пар кандидатов, имеющих дальние связи, нам нужно провести несколько шагов пере-инициализации параметров Expectation Maximization. Данная процедура может существенно изменить вероятности связей и даже сделать какие-либо из них несущественными для грамматики. Несколько шагов тренировки могут привести к дальнейшему отсеву мусора среди пар кандидатов. К сожалению, делать переобучение нужно не целиком в основной памяти компьютера (для больших словарей порядка 100 тыс слов её может не хватить), а через последовательную обработку файлов корпуса на жестком диске. Данный этап работы автором еще не завершен. Кроме этого, качественная статистическая модель обязательно содержит процедуры сглаживания, а это требует дополнительного программирования. Так как каждая пара приводит к уменьшению кросс-энтропии корпуса, то все пары в сумме также гарантировано должны приводить к снижению кросс-энтропии. Однако нам неизвестно, насколько велико будет суммарное улучшение и будет ли оно существенно лучше n-gram моделей. Проводить сравнение несглаженной дальнодействующей модели со сглаженной n-gram моделью не вполне корректно, так как несглаженные модели существенно хуже, чем сглаженные. После настройки параметров вероятностной модели, мы можем подключить нашу модель языка в какую-либо практическую систему для измерения качественных результатов. К примеру известно, что в системах распознавания речи, где также используются статистические модели языка, число ошибок линейно уменьшается в зависимости от кросс-энтропии. Мы также можем подключить модель языка к системе статистического машинного перевода и измерить улучшение по стандартной BLEU метрике, хотя у нас есть подозрения, что BLEU метрика не увидит улучшения, так как использует ngram-совпадения при сравнении переводов. Человеческие оценки качества перевода несколько затратны и не могут быть осуществлены для больших корпусов. Таким образом, за неимением лучшего, мы будем использовать кросс-энтропию на тестовом корпусе как самый главный критерий качества нашей языковой модели.

Список литературы

1. [Protasov 06] Протасов С . В . Обучение с нуля грамматики связей русского языка. //Десятая национальная конференция по искусственному интеллекту с международным участием., КИИ-2006.

2. [Lafferty et al. 92] Lafferty J. Sleator D. Temperley D. Grammatical Trigrams: A Probabilistic Model of Link Grammar. //Proceedings of the AAAI Conference on Probabilistic Approaches to Natural Language, 1992.

3. [Pietra et al. 94] Pietra S., Pietra D., Gillet J., Lafferty J., Prinz H., Ures L. Inference and Estimation of a Long-Range Trigram Model. //Grammatical Inference and Applications, Second International Colloquium, ICGI-94, 1994.

4. [Sleator and Temperley, 91] Sleator D. Temperley D. Parsing English with a Link Grammar.//Carnegie Mellon University Computer Science technical report CMU-CS-91-196, 1991.

5. [Jelinek, 97] Jelinek F. Statistical Methods for Speech Recognition. //MIT Press. ISBN: 0-262-10066-5. M.: 1997.

069-Protasov:_Layout 1 18.05.2008 16:05 Page 448

6. [Manning and Shutze, 99] Manning C., Schutze H. Foundations of Statistical Natual Language Processing. //Cambridge, MA: MIT Press.M.: 1999.

7. [Baum 72] Baum L. E. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. //Inequalities, 627(3):1-8,M.: 1972.

8. [Brown, 92] Brown P. F. Stepen A. L. An estimate of an upper bound for the entropy of English. //Computational Linguistics. 1992.

9. [Lari and Young, 90] Lari K. Young S. J. The estimation of stochastic context-free grammars using the inside-outside algorithm. //Computer Speech and Language. 1990.

070-Rozina_P:_Layout 1 18.05.2008 16:06 Page 449

НОМИНАЛИЗАЦИИ В РАЗГОВОРНОЙ РЕЧИ[1]
NOMINALIZATIONS IN EVERYDAY SPEECH

Розина Р.И. *(rarozina@yandex.ru)*
Институт русского языка им. В.В. Виноградова РАН

Статья посвящена сопоставлению номинализаций в русской разговорной речи, в современном сленге и в русском литературном языке. Рассматриваются источники номинализаций, модели их словообразования, семантика, модели управления и поверхностное поведение. Автор приходит к выводу о том, что, учитывая промежуточное положение номинализаций между глаголом и существительным, русские разговорные и сленговые номинализации отстоят от мотивирующих их глаголов дальше, чем номинализации в рамках русского литературного языка.

Высокочастотное употребление номинализаций (абстрактных отглагольных существительных) традиционно считается характеристикой официально-деловой речи, но в последнее время номинализации проникают в рекламу и в разговорную речь. Этот процесс противоположен проникновению сниженной (разговорной и сленговой) лексики в официальную речь, что также отмечается в последние десятилетия; но и в том, и в другом случае реализуется общая тенденция смешения стилей, характерная для современной речевой практики.

Наряду с «официальными» номинализациями (1) , в рекламе и разговорной речи встречаются отглагольные существительные, имеющие стилистически сниженный характер (разговорную или сленговую окраску) (2)-(6):

(1) Дивв. Пластиковые окна. Бесплатный *выезд* специалиста. Бесплатный *замер*. Бесплатная *доставка*. Короткие сроки *изготовления*. Только до 30-го июня 2004 г. при *оформлении* заказа «под ключ» (реклама фирмы «Дивв»).
(2) Цены смешные! Просто *улет*! Бренды *улет*! (реклама Мегамагазина электроники)
(3) К мужскому празднику подарки – *улет*! (реклама магазина Медиамаркт)
(4) Как всегда, вышел полный *облом* [‘ничего не получилось’](устная речь, дек. 2007).
(5) Это безобразие! Это просто *расшивка* [‘намеренный развал судебного дела’] ! (обманутая вкладчица, ТВ 13.12. 2007)
(6) Они это сделали по Татьяниной *наводке* (разговорная речь, февр. 2008)

Доклад посвящен особенностям разговорных и сленговых номинализаций, которые отличают их от номинализаций в русском литературном языке.

1. Источники номинализаций в разговорной речи

По своему происхождению разговорные и сленговые отглагольные существительные могут принадлежать одной из трех групп:

а) наиболее многочисленная группа - производные от разговорных или сленговых глаголов, например *балдеж* (разг.) ‘получение удовольствия от праздного времяпрепровождения’от *балдеть* (разг.) ‘получать удовольствие от безделья’; *обувалово* ‘ситуация отъема денег или вещей обманом или силой’от *обуть* (сленг) ‘ограбить, отнять’ (в буквальном и переносном смысле); *заказ*, *заказуха* 1 ‘убийство’ от *заказать* (сленг) ‘дать поручение убить кого-либо за плату’; *наводка* ‘совет, информация, указание’ от *навести* ‘указать, дать информацию’, *кидалово* ‘получение денег нечестным путем, без выполнения обязательств’ от *кидать* ‘получать деньги нечестным путем, не выполняя обязательств’. Лексические значения этих существительных не отличаются от значений мотивирующих их сленговых глаголов[2] – иными словами, они представляют собой синтаксические дериваты (Курилович: 1962).

[1] Статья выполнена в рамках проектов «Русская литературная норма и современная речевая практика (социолингвистическое исследование)« (проект выполняется в рамках программы ОИФН РАН «Русская культура в мировой истории») и проекта «Актантная структура глагола и отглагольного имени» (грант РГНФ № 08-04-001-81а).

[2] Возможно, что суффикс существительных усиливает отрицательную оценку, которая уже есть в семантике мотивирующего глагола, (ср., например, *кидать – кидалово*, *обувать - обувалово* и *заказать - заказуха*), но это несущественно меняет значения лексемы.

070-Rozina_P:_Layout 1 18.05.2008 16:06 Page 450

Иногда при номинализации сленговых глагольных лексем возникают лексемы, омонимичные лексемам существительных литературного языка, которые представляют собой номинализации этих же глаголов, но в других значениях, например *откат* - действие по глаголу *откатить* ‘катя, переместить в сторону на какое-то расстояние’ (МАС) и (сленг) ‘взятка чиновнику за предоставление заказа, сулящего большую прибыль’, *накат* ‘ряд бревен, досок, уложенных, настланных сверх чего-то или под чем-л.; настил’ (МАС) от *накатить* ‘катя, надвинуть на что-то, покрыть чем-л. какую-л. поверхность’ и (сленг) ‘нападки, угрозы’от *накатить* ‘приехать в каком-л. (обычно большом) количестве’, *отмывание* – действие по глаголу *отмывать* ‘мытьем удалять (грязь, пятна и т.п.)’(МАС) и (сленг) ‘превращение «грязных» денег в чистые путем проведения их через законные операции’ от соответствующего сленгового глагола[3]. В подобных случаях различие между сленговым и литературным существительными определяется только связью с мотивирующими их значениями глагола.

б) производные от глаголов литературного языка, например *заказуха* 2 (сленг) ‘произведение искусства, сделанное на заказ’ от лит. *заказать*; *стрелка* (сленг) ‘встреча бандитских группировок’, ‘выяснение отношений между бандитскими группировками, сопровождающееся применением огнестрельного оружия’ от *стрелять, сходняк* ‘собрание воров в законе’ от *сходиться*, *оживляж* ‘средство сделать что-то более занимательным’ от *оживить.* При образовании этих существительных, в отличие от существительных первой группы, лексическое значение производящего глагола модифицируется. Появляются семантические приращения, в результате которых значения этих существительных оказываются более специализированными и по сравнению с значением мотивирующего глагола, и по сравнению с значением существительного, образованного от того же глагола в литературном языке. Так, у номинализации глагола *заказать* в литературном языке – *заказ* – нет никаких семантических признаков, сужающих или расширяющих значение по сравнению с значением глагола, и, соответственно, она не предполагает никакого ограничения на синтаксический объект, ср. *заказать суп* / *машину* / *эскиз медали* / *сочинение композитору* / *хор*[4] – *заказ супа* / *машины* / *эскиз медали* / *сочинения композитору* / *хора*. Между тем сленговая номинализация этого же глагола – *заказуха* – обозначает результат действия *заказать* со специализированным объектом ‘произведение искусства’. Кроме того, у сленгового отглагольного существительного *заказуха* есть отрицательная коннотация, которая вносится суффиксом *–ух*(*а*), отсутствующая как у мотивирующего глагола, так и у литературной номинализации *заказ*. Ее появление связано как с оценкой результата действия – произведение, выполненное на заказ, оценивается отрицательно, так и с характером суффикса.

Образованию сленгового отглагольного существительного может сопутствовать опустошение значения вплоть до того, что в нем остается один семантический признак – оценка, что никогда не имеет места при номинализации в рамках литературного языка или в терминологии. Так, в литературном языке от глагола *отпасть* в значении 2 ‘утратить связь, выйти из состава какого-н. объединения’ образовано отглагольное существительное отпад с соответствующим значением например:

(7) Но если будет продолжаться такая политика, то *отпад* некоторых национальных окраин от России возможен (Беседа на радио «Эхо Москвы» 5.09.2006).

Кроме того, от этого же глагола в другом значении – ‘оторваться. Отвалиться, отделиться, падая’ - образован употребительный и в настоящее время специальный термин лесоводства *отпад*, например:

(8) К сожалению, процент лесного *отпада* из года в год возрастает, в результате чего увеличиваются очаги отмерших деревьев, и древостои начинают распадаться (Лесное хозяйство, 2004 (НКРЯ)).

Омонимичное сленговое существительное *отпад* – предикативная метафора, выражающая высшую положительную оценку, например:

(9) Назаров играет потрясающе. Его дуэт со Степановым в «Гражданине начальнике» это я не знаю, как нынче говорят полный *отпад* (звонок слушателя на радио «Эхо Москвы» 25.02.2006).

в) существительные, образованные от глаголов литературного языка по аналогии с уже существующими номинализациями сленговых глаголов, например *накат* от *накатить* по аналогии с *наезд* от *наехать* (сленг) ‘атаковать’, ‘налететь с угрозами или упреками’ при том, что у *накатить* нет значения, синонимичного *наехать*, из чего следует, что сленговое значение слова *накат* нельзя вывести из значения производящего глагола; *обиралово* от *обирать* и *обдиралово* от *обдирать* по аналогии со сленговым *обувалово*.

2. Словообразовательные модели

Отглагольные существительные в разговорной речи образуются либо с помощью суффиксации (суффиксы *–ух(а)*, *-ов(о)*, *-овк(а)*, *-к(а)*, *-д*, *-яж*, *-нщ*(*е*) и др.), либо усечением (*залет*, *прокол*, *улет*). Особенность сленговой номинализации – вариативность производных одного и того же глагола, например *мочить* – *мочиловка* и

[3] Между литературными и сленговыми лексемами существительных в перечисленных парах – не чистая омонимия, потому что между ними можно усмотреть отдаленное семантическое сходство.

[4] Источник примеров – Национальный корпус русского языка (НКРЯ)

070-Rozina_P:_Layout 1 18.05.2008 16:06 Page 451

мочилово; *отмывать* - *отмыв*, *отмывка* и *отмывание*; *расслабуха* и *расслабон* – номинализации глагола *расслабляться*.

Сравнение номинализаций одного и того же глагола в сленге и в литературном языке позволяет прийти к выводу, что часто, хотя и не всегда, в сленге выбирается иной способ номинализации, чем в литературном языке. Прежде всего, существуют случаи, когда в сленге номинализация есть, а в литературном языке она невозможна, например *отрубиться* (сленг) 'перестать воспринимать реальность' - *отруб*; *улететь – улет*. Там, где литературный язык прибегает к суффиксации, сленг выбирает усечение, например *напряжение* в литературном языке и *напряг* в сленге. Там, где в литературном языке используется один суффикс, сленг выбирает другой, например номинализация глагола *объявить* в литературном языке – *объявление*, а в сленге - *объява* .

В сленге используются тот же набор суффиксов, что и в литературном языке, но не соблюдаются семантические ограничения на тип основы, к которой они применяются. Так, суффикс *-ух(а)* в литературном языке используется для образования наименований лиц женского пола от глаголов по типичной деятельности (*стряпать– стряпуха*) и от прилагательных по свойствам (*старый – старуха*), а также животных (*белуха*, *лысуха*), растений (*синюха* , *зеленуха*) и болезней (*краснуха*, *золотуха*) по их типичным признакам[5]. В разговорной речи и сленге суффикс *–ух*(*а*) используется для образования названий состояний от глаголов (*расслабуха*, *показуха*) и для оформления усечений существительных (*порнография – порнуха*, *черный – чернуха, жизнь - житуха*).

В последнее время при образовании отглагольных сленговых существительных как от литературных, так и от сленговых глаголов стал продуктивным суффикс

-ов(*о*), вытеснивший продуктивный ранее суффикс *-овк*(*а*), ср. *кидалово* 'ситуация обмана, часто связанная с тем, что человек тратит деньги против своего желания', *махалово* 'драка', *обувалово* 'ситуация, в которой человек против своей воли тратит большое количество денег', *обдиралово*, *обиралово*. В литературном языке суффикс

-ово.и его варианты *–ив*(*о*) и *-ев*(*о*) используются для образования отглагольных существительных, обозначающих вещества или массы по действию, которое производится над ними *курево*, *топливо*, *хлёбово*,или в результате которого они возникают, ср. *месиво*, *печево*.

Усечение при образовании сленговых номинализаций используется во многих случаях тогда, когда в литературном языке отглагольное существительное от этого же глагола образуется с помощью суффиксации, ср. *напряг* в сленге и *напряжение* в литературном языке от *напрягать*,

3. Таксономические категории

Среди разговорных и сленговых номинализаций встречаются существительные категории СОСТОЯНИЕ (*балдеж, отпад*, *расслабон*), ДЕЯТЕЛЬНОСТЬ (*гудеж*, *зачистка, разборка*), ДЕЙСТВИЕ (*наезд*, *накат*, *подстава*), ПРОИСШЕСТВИЕ (*залет*, *облом*, *прокол*). Отсутствуют существительные категории ПРОЦЕСС, что соответствует отсутствию сленговых глаголов этой категории.

4. Модели управления

Известно, что при номинализации глагола число участников ситуации остается прежним, но совсем не все из них получают поверхностное выражение. Так, при номинализации переходных двухвалентных глаголов поверхностное выражение получает только один их двух участников, причем, как правило, тот, который является вторым актантом глагола (Dik 1997: 158), т.е. его синтаксическим объектом, например:

(10) а. *Микеланджело создал скульптуру Давида* за два года.
б. *Создание скульптуры Давида* заняло у Микеланджело два года.

В этом номинализация сходна с пассивизацией: первый актант (Агенс) теряет выражение (уходит за кадр) или понижается в ранге и становится сирконстантом:

(11) а. *Микеланджело создал скульптуру Давида* за два года.
б. *Скульптура Давида была создана* (Микеланджело) за два года.

При номинализации глаголов в разговорной речи актантные преобразования идут дальше: как правило, поверхностного выражения не получает ни один участник, например:

(12) Многие из нас учатся… Этому посвящена тема. Тому, как *нас напрягает учеба* (http://www.cheat-world.ru/forum/showthread.php?t=26221).
Если ты выбрал профессию журналиста, то на него можно *учиться без напряга*, с удовольствием (http://gtfo.ru/2007/11/10/ya-normalnyj/)

[5] Грамматика-80: 152, 174.

070-Rozina_P:_Layout 1 18.05.2008 16:06 Page 452

Невозможность поверхностной реализации участников ситуации при номинализации в разговорной речи может быть вызвана несколькими причинами:

а) Пациенс был инкорпорирован ДО номинализации производящим глаголом.

У различных лексем глагола *поддать* в пределах литературного языка

объектная валентность - переменная, ср. *поддать мяч/ попону ногой*, *поддать пару.* У сленговой лексемы *поддать* объектный актант 'алкоголь' инкорпорирован и наследуется отглагольным существительным *поддача*:

(13) а. Они *поддают*.

б. Я представил, как он после *поддачи* перед банькой, озираясь дает Брежневу почитать эти стихи. (Андрей Вознесенский. На виртуальном ветру (1998). (НКРЯ))

б) Пациенс инкорпорируется одновременно с номинализацией глагола, ср.модели управления глагола *предъявлять* в официальной речи, где инкорпорации не происходит:

(14) а. *Покупатели предъявляют претензии* магазину в письменной форме.

б. *Предъявление претензий* должно осуществляться в письменной форме.

и в разговорной речи:

(15) Вот / начал там такие *предъявы* типа кидать [Реалити-шоу «Дом-2» (2006.04)] (НКРЯ).

в) При заполнении валентности возникает неоднозначность, ср.:

(16) а. Бандитам *заказали Пугачеву*. Угонщики выслеживали новый джип примадонны несколько недель (День звезды. 7.01.2008).

б. *Заказ* Пугачевой ['заказ, который сделала Пугачева' или 'заказ что-то сделать с Пугачевой'].

г) При образовании сленговой лексемы от глагола происходит смена таксономической категории: от глаголов действия, процесса или происшествия образуются существительные категории состояния, которые по определению не могут иметь объект, ср. номинализации в литературном языке (17а) и (18а) и в сленге (17б) и (18б):

(17) а. автономная область *отпала* (происшествие) - *отпад* (происшествие) автономной области

б. Какой концерт! Просто *отпад*! (состояние)

(18) а. *напрячь* мышцы – *напряжение* мышц

б. - Ржевский помялся и закончил: - Все нормально и без *напрягов* (Леонов, Макеев. Эхо дефолта (2000-2004) (НКРЯ).

Отглагольные существительные – часть речи, занимающая промежуточное положение между глаголами и существительными. В разных языках, благодаря особенностям своего поверхностного синтаксиса, они оказываются ближе к тому или другому полюсу (Koptjevskaja-Tamm 1993: 6-7).

Наш материал показывает, что и в разных социолектах позиция отглагольных существительных по отношению к глаголу и существительному может быть различной.

В русской разговорной речи и сленге отглагольные существительные с их тенденцией к полному отказу от поверхностного выражения актантов ближе к существительным, чем номинализации в рамках литературного языка, модель управления которых допускает поверхностное выражение хотя бы одного из актантов мотивирующего глагола.

Список литературы

1. Грамматика-80. Русская грамматика. Т.1. М.:Наука, 1980.

2. Курилович Е. Деривация лексическая и синтаксическая // Курилович Е. Очерки по лингвистике. М., 1962.

3. Dik S.C. The theory of Functional Grammar. Part II. Complex and derived constructions. (Second revised edition). Ed. Kees Hengeveld // Berlin; New York: De Gruyter, 1997.

4. Koptjevskaja-Tamm M. Nominalizations. L.; N.Y., 1993.

071-Rubashkin_VSh:_Layout 1 18.05.2008 16:12 Page 453

ОНТОРЕДАКТОР КАК КОМПЛЕКСНЫЙ ИНСТРУМЕНТ ОНТОЛОГИЧЕСКОЙ ИНЖЕНЕРИИ[1]

ONTOLOGY EDITOR AS INTEGRATED DEVELOPMENT ENVIRONMENT

Рубашкин В.Ш. *(VRubashkin@yandex.ru),* ***Пивоварова Л.М.*** *(pivovarova@iphil.ru)*
Санкт-Петербургский государственный университет

В докладе представлен опыт разработки и использования онторедактора, ориентированного на модель знаний онтологии InTez. Рассматриваются функции просмотра, ввода и редактирования, тестирования и др. Проводится сопоставление с зарубежным опытом аналогичных разработок.

Онторедакторы представляют сравнительно новый вид информационных технологий; требования к ним и представления об их функциональности еще только формируются. Разрабатываемые средства для работы с онтологиями весьма разнородны: они могут быть ориентированы на определенную модель знаний; иметь многомодульную или интегрированную архитектуру; поддерживать тот или иной набор функций, использовать разные методы и технологии. Несомненно, однако, что критическая масса результатов уже налицо;[2] движение в сторону унификации, как и в сторону объединения разных по назначению инструментов в интегрированный комплексный продукт достаточно хорошо различимо. Цель настоящего доклада – попытаться обозначить общие тенденции и сформулировать некоторую, как мы надеемся, последовательную концепцию построения такого рода инструментов. При этом авторы опираются – не в последнюю очередь – и на собственный опыт (онторедактор ***InTez***, ориентированный на модель знаний, подробно описанную нами в работе [5]) и иллюстрируют им возможность предлагаемых решений. Доклад не является обзором конкретных онторедакторов – существующие обзоры (см. [1, Ch 5], [2, Part II], [3]) в своей совокупности дают достаточно полную картину сложившегося в этой области исследований и разработок положения.

1. Общие замечания

Границы понятия «онтология» разными авторами проводятся по-разному (ср., напр., [1, Ch 5]). Не имея возможности здесь входить в обсуждение этого вопроса, обозначим коротко то понимание, которое далее будет иметься в виду.[3]

1. Единицей описания в онтологии является **понятие**; термин *концепт* мы будем употреблять просто как его синоним. Понятия формируются в сфере профессиональной деятельности; это понимание систематически фиксируется в учебниках, обзорах и энциклопедических словарях. И **в этом смысле** любая онтология есть формальная модель лексической системы профессионального языка.

2. Онтология базируется на некоторой **модели знаний**. Под моделью знаний мы понимаем язык представления знаний (ЯПЗ) вместе с некоторым набором схем аксиом, определяющих возможности системы вывода. Специфика модели знаний предполагает соответствующую специализацию инструментальных средств. Формализация означает, в частности, представление единиц практически сложившихся терминосистем как конструктов выбранного ЯПЗ, в идеале – как логических формул некоторого логического исчисления, на которое с самого начала накладываются ограничения, касающиеся как выразительных возможностей, так и допустимых схем логического вывода. Установка на интеграцию в едином онтологическом представлении разнородных профессиональных систем предполагает обязательное наличие интегрирующей их онтологии «верхнего уровня» (Top-Level). Это, а также сама необходимость организации эмпирически данного лексического материала в соответствии с требованиями принятой модели знаний делает необходимым введение в онтологию значительного числа «фиктивных» (изобретаемых разработчиком онтологии для соблюдения условий целостно-

[1] Работа выполнена при финансовой поддержке РФФИ (проект № 06-06-80434)

[2] В обзоре [3], например, дано краткое описание 93-х онторедакторов.

[3] Более подробно наше понимание общих вопросов онтологического моделирования изложено в докладе, представленном в материалах предыдущей конференции «Диалог'2007» ([4]).

071-Rubashkin_VSh:_Layout 1 18.05.2008 16:12 Page 454

сти) концептов. Онтология всегда предполагает определенный компромисс между «чистотой» модели знаний и практической мотивированностью реально сложившейся терминологией, определяющей – вместе с набором «фиктивных» терминов и некоторым количеством вовлекаемых в процесс профессиональной коммуникации слов повседневного языка – способы именования концептов.

3. Онтология обладает вычислительной функциональностью. Можно считать, что эта функциональность воплощена в онтологическом API, реализующем некоторый доступный любым приложениям набор программных функций. Среди них обязательно присутствуют функции, реализующие процедуры ограниченного логического вывода.

Многие характеристики онторедактора, включая его функциональные возможности, существенно зависят от базовой модели знаний, принимаемой в том классе онтологий, на который редактор ориентирован. В определенной степени характеристики и возможности онторедактора зависят также от используемой операционной среды (СУБД, XML, текстовый процессор и др.) и диктуемого ею представления данных. В частности, выбор, например, в онторедакторе ***InTez*** в качестве базовой операционной среды реляционной СУБД и естественной для нее SQL-техники манипулирования данными практически предопределяют способ реализации таких функций как поиск и формирование выборок; реализация других функций существенно опирается на возможности такой операционной среды.

Таким образом, можно сказать, что онтология (понимаемая как информационно-вычислительный ресурс) предоставляет **программный интерфейс** приложениям; онторедактор реализует **человеко-машинный интерфейс**, обеспечивающий администрирование онтологий; для реализации части функций онторедактора должна использоваться функциональность самой онтологии.

2. Функциональность онторедактора

Вполне очевидны такие функции как навигация, броузинг и поиск; ввод и редактирование. В силу специфики инструмента к ним прибавляются другие, в том или ином виде реализуемые в разных проектах: поддержка (или использование) машины ограниченного вывода (*reasoner, inference engine*); средства тестирования онтологии.

Есть и другие аспекты, так или иначе определяющие характер функционирования онторедактора:

• организация взаимодействия с пользователем (включая наличие графического интерфейса);

• возможности и средства доменного редактирования и интеграции разнородных концептуальных систем; основным условием здесь является наличие встроенной онтологии верхнего уровня (*Top-Level Ontology*), без которой, как нам представляется, онторедактор теряет способность объединять и интегрировать концептуальные модели разных предметных / проблемных областей;

• средства и способы представления экземпляров, являющихся «примерами» (instance) концептов онтологии; способы работы с «описаниями экземпляров».

3. Навигация, броузинг и поиск

Специфика онторедактора такова, что даже эти вполне традиционные для любого редактора функции требуют обсуждения. Просмотр и навигация предполагают, прежде всего, некоторую «естественную» упорядоченность материала. В текстовом редакторе смысл этого выражения вполне очевиден – это порядок следования слов и предложений в тексте. Применительно к онторедактору, мы, возможно, склонимся к выводу, что естественного порядка в концептуальной модели вообще не существует. Действительно, мы можем говорить о физическом порядке следования записей, об упорядоченности по ключу, или об алфавитном порядке терминов, но все это, с точки зрения концептуальной модели, не касается существа представляемого онторедактором материала. Алфавитный порядок имеет значение, но, скорее, как поисковый индекс, обеспечивающий быстрый поиск нужного пользователю термина. В такой ситуации поиск, формирование выборок (фильтры) и навигация по связям разного типа оказывается существенной поддержкой для реализации комфортного взаимодействия пользователя с концептуальной системой. «Естественной» для концептуальной системы можно считать, скорее, таксономическую (*общее - частное*) упорядоченность концептов; она образует ядро всякой концептуальной модели. Так что «естественным» порядком просмотра и навигации здесь скорее является просмотр «сверху вниз» (от общего к частному). А также, возможно, просмотр групп концептов связанных иерархическими связями другого типа (например, *целое - часть*). Но иерархическая упорядоченность не является линейной, и уже одно это порождает совсем другие требования к интерфейсу. В частности, возникает потребность графического представления всех или некоторых связей между концептами и поддержки процедур графического редактирования, - что становится стандартом де-факто для такого рода инструментов.

Помимо перечисленного, к этой группе функций следует отнести операции, позволяющие устанавливать

071-Rubashkin_VSh:_Layout 1 18.05.2008 16:12 Page 455

взаимное соответствие концептов и единиц ЕЯ: получение множества слов / словосочетаний, выражающих данное понятие (список синонимов), либо множества концептов, которые может выражать отдельно или совместно с другими словами данное слово – омонимия.

Возможности онторедактора ***InTez*** по этой группе функций можно дополнительно охарактеризовать следующими замечаниями. В главное окно онторедактора (см. рис. 1) всегда выводится дерево признаков (аналог таксономической иерархии в используемой модели знаний). Опционально выводятся таблицы, представляющие общий список дескрипторов, и набор специализированных вкладок. Дополнительно может быть выведено окно словарной статьи текущего или вновь вводимого концепта, где наименования и значения словарных характеристик представлены в вербальной форме. Имеется возможность поиска термина и синонимов по строке-образцу (с использованием простейших регулярных выражений) – с последующим последовательным просмотром всех найденных, а также возможность поиска концепта по ключу. В строке состояния при этом отображается число найденных по образцу концептов.

Графический интерфейс реализуется посредством стандартного объекта ***TreeView***, соответственно, пользователю доступна вся его функциональность – как в отношении просмотра, так и в отношении графического редактирования. При этом обеспечивается графическое различение разных категорий концептов, отображаемых в дереве признаков (*классификационные, количественные, строковые признаки, наименования групп признаков, базовые свойства*). Для удобства просмотра к стандартной функциональности ***TreeView*** добавлена опция «Показать куст», позволяющая компактно вывести на экран всех «братьев» (*sibling nodes*) указанного узла.

На альтернативных вкладках можно просматривать и редактировать:

- связи текущего концепта;
- представляющие его слова и словосочетания («Лексикон»);
- тексты определений.

При просмотре связей пользователь имеет возможность выбрать просмотр исходящих либо входящих связей, а также осуществлять переход по связям к любому из ассоциированных концептов. При работе с Лексиконом доступны обе указанные выше опции («синонимия» и «омонимия»).

При необходимости могут быть установлены фильтры, ограничивающие доступ и просмотр только концептами определенной категории и типа (семантический класс - подкласс), а также фильтр по параметрам администрирования (имя администратора и/или дата редактирования).

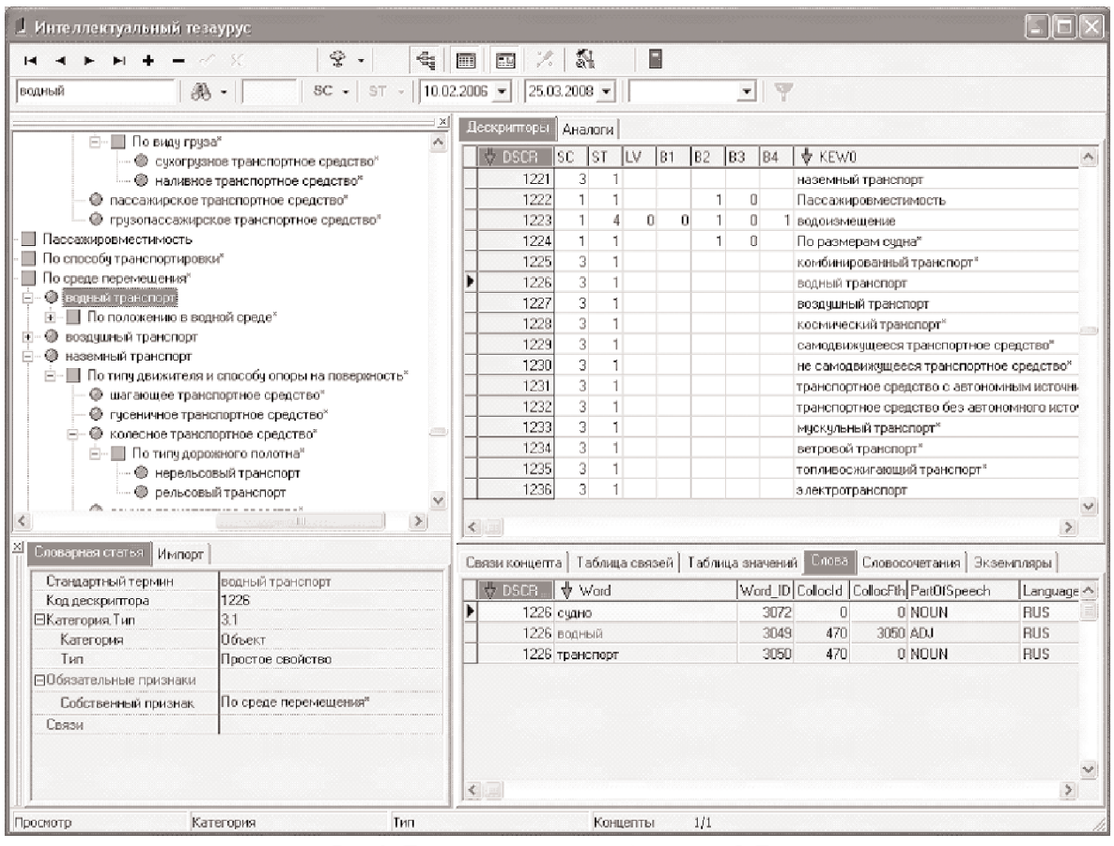


Рис.1 Главное окно онторедактора InTez

071-Rubashkin_VSh:_Layout 1 18.05.2008 16:12 Page 456

4. Ввод и редактирование

Словарная статья концепта всегда включает **унарные характеристики** концепта и **отношения**, связывающие его с другими концептами. При этом нужно различать 2 типа отношений:

а) **Бинарные отношения** концепт – концепт. На языке логики они суть не что иное как представление логических постулатов значения на языке описания словарных статей. (*«тигр - хищник»; «травоядное – не хищник»; «всякое животное имеет голову»; «*условие применимости признака *должность* – концепт *работающий по найму»* и т.п.) Следует различить два типа таких отношений: объемные (*включение, совместимость, несовместимость*) и все прочие; последние – по аналогии с терминологией, сложившейся в практике разработки традиционных информационно-поисковых тезаурусов – можно именовать *ассоциативными.*

б) В онтологии ***InTez*** представлены также **дефиниционные отношения** – отношения между определяемым концептом и концептами, входящими в формальное толкование определяемого (предварительно должны быть специфицированы допустимые схемы формальных толкований).

Кроме того, словарные характеристики могут быть обязательными либо необязательными, повторяющимися либо уникальными.

В современной онтологической инженерии рассматриваются три возможных способа пополнения онтологий:

a) «ручной» ввод;

b) автоматический или автоматизированный ввод на основе анализа корпуса текстов;

c) автоматический или автоматизированный ввод с использованием традиционной лексикографической информации (энциклопедических и толковых словарей).

Поскольку варианты, указанные в п.п. b) и c) фактически представляют собой высокоспециализированные и не достигшие еще достаточной зрелости технологии,[4] здесь мы будем обсуждать только процедуры ввода в смысле п. a) - ввод в собственном смысле слова.

Онтология, рассматриваемая как информационно-вычислительный ресурс для поддержки широкого спектра интеллектуальных информационных технологий, предъявляет жесткие требования к достоверности ввода. Это, собственно говоря, и есть основная проблема, которая должна решаться при проектировании процедур ввода. Другая актуальная в любых системах ввода и редактирования проблема – проблема эргономичности – здесь тесно связана с первой, и способ и качество решения второй в значительной степени зависит от способов и качества решения первой.

Требование достоверности ввода может быть конкретизировано в следующих пунктах

1) Неизбыточность и полнота описания – должны быть определены те и только те словарные признаки, которые релевантны для концептов данного типа.

2) Непротиворечивость описания – словарные характеристики не должны противоречить друг другу. Скажем, для концепта, определяемого конъюнкцией объектных классов (в терминах онтологии ***InTez*** - *И-толкование*; в терминах многих других онтологий – класс, характеризуемый через множественное наследование), определяющие концепты должны быть *совместимы* (в терминах ***OWL*** – не должны находиться в отношении *Disjoint*). Так что процедура ввода должна обнаруживать и блокировать ввод , например, И-толкования вида *X ≡ животное And металлический.*

3) Правильность означивания – вводимые значения определяемых словарных признаков должны принадлежать области их допустимых значений. Скажем, формально неправильным будет указание в качестве базового признака для единицы измерения *метр* концепта *перемещение* (имеем легко контролируемую категориальную ошибку – базовым признаком может быть только концепт класса *сочетающийся с числом;* правильно будет *линейный размер*). Однако ошибка, состоящая в указании в той же ситуации в качестве базового признака концепта *масса,* уже не является формально контролируемой и может оставаться не выявленной до тех пор, пока онтология не начнет использоваться в приложениях, для которых именно эта связь окажется существенной. Если допустить, например, что концепт *лед* администратор пытается определить как конъюнкцию концептов *агрегатное состояние* и *химический состав*, будет обнаружена формальная ошибка, состоящая в том, что формальное толкование типа «конъюнкция» для объектного термина может содержать только объектные термины, либо означенные признаки. Однако определение типа *лед ≡ квазиобъект And цилиндрической формы* уже не содержит ошибок такого рода и является формально правильным.

4) Содержательная правильность – вводимые словарные характеристики должны быть адекватны смыслу добавляемого или редактируемого концепта. (Скажем, ошибкой этого типа будет отнесение администратором признака *'цвет'* к группе *химические свойства вещества;* такого же рода ошибки демонстрируют примеры п. 3)).

---

[4] В англоязычной литературе они объединяются термином ontology learning (ср. [1, § 3.5]).

071-Rubashkin_VSh:_Layout 1 18.05.2008 16:12 Page 457

Конечная цель при проектировании процедур ввода состоит в том, чтобы **полностью исключить** формально определимые ошибки, т.е. ошибки, соответствующие п.п. 1), 2) и 3). При этом технологически «хорошее» решение будет состоять не в том, чтобы уметь обнаруживать формальные ошибки *post factum*, а в том, чтобы сама процедура ввода была спроектирована так, что ввод логически некорректных элементов описания оказывается вообще невозможным. Это означает, что функцию контроля формальной корректности словарных описаний мы полагаем правильным переместить из подсистемы тестирования, куда она помещается сейчас большинством разработчиков онторедакторов[5], в подсистему ввода.

Что касается содержательных ошибок, то они могут возникать в силу случайной описки или неверно выполненного действия администратора, либо вследствие неполного или неправильного понимания им смысла вводимого концепта, как и концептов отнесения, связи с которыми фиксируются при определении вводимого концепта. Выявление такого рода ошибок представляет сложную и вряд ли окончательно и полностью разрешимую проблему для службы администрирования онтологии. Эта задача находится в компетенции подсистемы тестирования онтологии.

Существенно, что решение задач формального контроля обусловлено возможностью построить формальное описание системы словарных признаков. Таковое сводится к определению области значений каждого признака и к установлению отношений зависимости по условиям применимости между признаками. С точки зрения первого требования признаки можно разделить на признаки со стандартной областью значения (вещественные, целочисленные, строковые) – здесь процедура формального контроля значения тривиальна, - и признаки, областью значений которых является некоторый класс концептов. Здесь важен выбор адекватной данной задаче схемы категоризации концептов.

Определение связей по условиям применимости в онтологии ***InTez*** соответствует схеме аксиом вида

$\forall x\ (\exists\ v\ (P\ (x, v)\ \leftrightarrow U\ (x))$,

где P – некоторый словарный признак, U – концепт, представляющий условие его применимости. Если принять в онтологии ограничение, что U – всегда есть значение некоторого другого по отношению к P классификационного признака, получаем структуру типа «дерево признаков» [5], для которой перебор всех релевантных признаков представляет собой алгоритмически простую задачу обхода дерева с двумя типами вершин. При этом система словарных признаков может быть включена в саму онтологию (добавлением узла *концепт* и поддерева конкретизирующих это понятие классификационных признаков). Таким образом онтология наделяется способностью **самоописания**.

5. Тестирование

Проверить содержательную правильность описаний концептов в рамках технологии администрирования онтологии – помимо прямого просмотра словарных статей – можно только путем организации «лабораторных» испытаний и экспертной оценки их результатов администратором. Понятно, что окончательную проверку и отладку («бета-тестирование») онтология может пройти в рамках целевых информационных технологий, скажем, в процедурах ее использования в системах анализа текста.

«Тестирование» отдельных концептов сводится к просмотру и проверке содержимого словарных статей и, следовательно, относится к компетенции подсистемы навигации и броузинга. Собственно тестирование как отличающаяся от броузинга процедура может состоять только в тестировании **отношений** между концептами – как объемных, так и ассоциативных.

В онторедакторе ***InTez*** тестирование выполняется в отдельном окне (рис. 2); результаты тестирования представляются в графической (представление объемных отношений диаграммами Венна) и текстовой (представление ассоциативных связей для тестируемой пары концептов) формах. Интерфейс позволяет эксперту оценить результат тестирования, так сказать, «одним взглядом». Возможно как «точечное» тестирование конкретной пары концептов, указываемых с использованием средств навигации, так и «серийное» тестирование, при котором пары концептов выбираются из онтологии случайным образом. Последний режим позволяет производить поиск ошибок путем быстрого «листания» произвольно выбираемых пар концептов. Поскольку исполнительная система онтологии интегрирована с онторедактором, нет необходимости для запуска процедуры тестирования дополнительно загружать и инициировать соответствующий модуль.

[5] Ср. напр. [6, p 50]: «…*standard service that is offered by reasoners is consistency checking*».

071-Rubashkin_VSh:_Layout 1 18.05.2008 16:12 Page 458

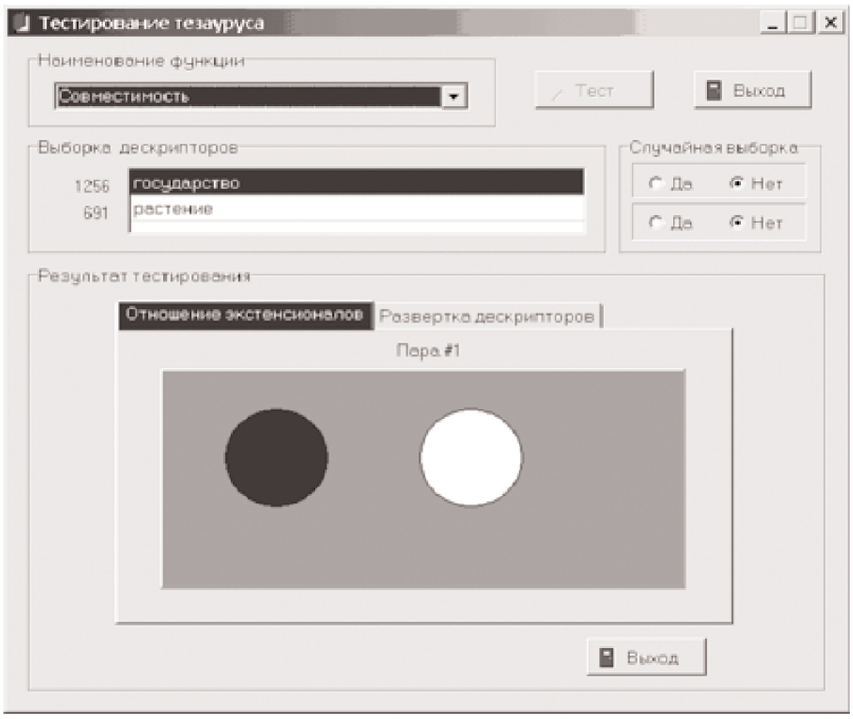


Рис. 2 Окно тестирования онтологии

6. Представление и работа с экземплярами

Во многих онторедакторах *экземпляры* (*индивиды*), во-первых, категориально отграничиваются от *концептов*[6] и, во-вторых, описываются другим набором словарных характеристик. Для этого можно найти основания в самой модели знаний: *общие понятия* характеризуются **применимостью / неприменимостью** признаков («атрибутов»), тогда как *экземпляры* требуют указания **значений** применимых к соответствующему общему концепту признаков. В самой общей форме: в терминах общих понятий формулируются законы, в терминах экземпляров – факты. Но с другой стороны, экземпляры являются конкретизацией соответствующих им общих понятий и наследуют всю относящуюся к последним словарную информацию. (Скажем, *Бразилия* унаследует от концепта *страна* информацию о том, что это понятие есть подкласс понятий *регион* и *социальный субъект*, характеризуется *численностью населения* и *размером территории*, имеет *столицу* и т.д.). На этом основании – и именно такое решение принято в онтологии ***InTez*** – экземпляры могут быть на общих основаниях включены в иерархию классов (с сохранением информации об «экземплярности»), будучи представлены в ней терминальными узлами.[7] Для удобства просмотра дерева имеется опция, позволяющая показать или скрыть все экземпляры.

В дальнейшем планируется интеграция онторедактора и поддерживаемой им онтологии с системой реляционных БД, в которых фактографическая информация об экземплярах должна храниться обычным образом в табличных записях. Имеется в виду, что при этом поддерживается связь между схемой БД (имена таблиц и полей, межтабличные связи) и онтологией таким образом, что все элементы схемы получают **концептуальную интерпретацию**. Таким способом предполагается придать онторедактору и поддерживаемой им онтологии функции интеллектуального интерфейса баз данных, обеспечивающего *прозрачный для смысла* (*sense transparent*) вербальный доступ к БД.[8]

[6] Термин *концепт* при этом употребляется как синоним выражения *общее понятие*.
[7] Не все терминальные узлы являются экземплярными.
[8] Одна из ранних попыток такого рода интеграции описана в [7, §7.2].

071-Rubashkin_VSh:_Layout 1 18.05.2008 16:12 Page 459

Список литературы

1. Gomez-Perez A., Fernando-Lopez M., Corcho O. Ontology Engineering // Springer–Verlag, 2004.

2. Staab S., Studer R. (eds). Handbook on Ontologies. // Springer—Verlag, 2004.

3. Denny M. Ontology Tools Survey, Revisited // [Электронный ресурс]: http://www.xml.com/pub/a/2004/07/14/onto.html - 2004.

4. Рубашкин В. Ш. Онтологии – концептуальные границы, проблемы и решения. Точка зрения разработчика // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007». М.: Издательский центр РГГУ, 2007. С. 481 – 485.

5. Рубашкин В. Ш. Представление и анализ смысла в интеллектуальных информационных системах // М.: Наука, 1989.

6. Horridge M. et al. A Practical Guide To Building OWL Ontologies Using The Protege-OWL Plugin and CO-ODE Tools Edition 1.0 // University Of Manchester, 2004.

7. Nirenburg S., Raskin V. Ontological Semantics // Cambridge, MA: MIT Press, 2004.

072-Ryko_A:_Layout 1 18.05.2008 16:13 Page 460

МНОГОУРОВНЕВАЯ ЛИНГВИСТИЧЕСКАЯ РАЗМЕТКА ЗВУКОВОГО КОРПУСА РУССКОГО ЯЗЫКА

MULTILEVEL LINGUISTIC ANNOTATION OF THE RUSSIAN SPEECH CORPUS

Рыко А.И. *(aryko@mail.ru),* ***Степанова С.Б.*** *(stsvet_2002@mail.ru)*
Лаборатория экспериментальной фонетики Института филологических исследований Санкт-Петербургского государственного университета

В докладе описываются возможности многоуровневой лингвистической разметки звукового корпуса русского языка для описания спонтанной речи и ее сопоставления с кодифицированным литературным языком.

Лингвистическое описание корпуса звучащих текстов предполагает осуществление его многоуровневой аннотации (разметки), первая и главная задача которой — установление инвентаря единиц этой разметки на разных языковых уровнях. В перспективе в рамках проекта «Интегральное моделирование звуковой формы естественного языка» предполагается разработка в той или иной степени автоматического аннотирования звуковых корпусов любого объема, однако на данном этапе мы ограничиваем свои задачи созданием инвентаря лингвистических единиц для такой разметки.

Как хорошо известно, привычные понятия фонемы, морфемы, слова и предложения оказываются неприложимыми или плохо приложимыми к спонтанной речи. Фактически на этом материале все традиционные метапонятия языка (единицы его описания) так или иначе разрушаются, на их месте создается нечто новое, что не всегда легко поддается определению и описанию (см.: Богданова 2006: 189–194).

Наша многоуровневая лингвистическая разметка как раз и является попыткой вычленить в звучащем (спонтанном) тексте все традиционные единицы лингвистического анализа — синтагмы, слова, фонемы, морфемы, а также слоги — и установить специфику этих единиц в спонтанном тексте.

В качестве материала для экспериментов по многоуровневой лингвистической разметке был взят «корпус медиков» - звуковая база данных, в которой представлены записи речи женщин-медиков (см.: Бродт 2007). В качестве «стартовой» единицы мы использовали синтаксические единицы («фразы»), соотносимые с традиционным предложением, которые были получены в результате проведенного предварительно синтаксического пунктирования текстов (см.: Там же). Всего было размечено 38 фраз (271 синтагма, 774 слова, 4140 звуков).

Многоуровневая лингвистическая разметка осуществлялась с помощью программы Praat, созданной сотрудниками факультета фонетики Амстердамского университета П. Бёрсма и Д. Вининком (Paul Boersma, David Weenink)[1], которая предназначена для фонетистов, исследующих звучащую речь.

Уровни лингвистической разметки

При аннотировании корпуса звучащих текстов мы выделяли единицы следующих лингвистических уровней.

1) Орфографический (Orthography).

«Рабочий» уровень, содержащий орфографическую запись (точнее, транслитерацию латиницей) анализируемой фразы.

2–4) Уровень пауз (Pause) и синтагм (Syntagma Ideal; Syntagma Real).

Согласно классическим определениям, *синтагма* — «фонетическое единство, выражающее единое смысловое целое в процессе речи-мысли» (Щерба 1955: 87), это «цепочка слов, связанная единым интонационным рисунком; членение потока речи на синтагмы достигается, с одной стороны, интонационным объединением внутри синтагмы, с другой — особым интонационным оформлением границы синтагмы»

[1] Создатели Praat регулярно обновляют версии программы и предлагают бесплатное использование ее для некоммерческих целей. Найти программу для бесплатного копирования можно на сайте www.praat.org.

072-Ryko_A:_Layout 1 18.05.2008 16:13 Page 461

(Гордина 1973: 211); «отрезок потока речи, заключенный между двумя паузами и характеризующийся усилением ударения на ударном слоге наиболее важного по смыслу слова (т. е. синтагматическим ударением) и объединяющей мелодикой» (Кравченко и др. 1973: 61–62)[2].

Устанавливая границы между единицами уровня Syntagma Ideal, мы, работая практически с орфографическим представлением устного текста, как бы прочитывали текст, ставя знаки границ там, где сами могли бы сделать паузу, при этом мы, в первую очередь, руководствовались смысловым единством внутри каждой единицы. Понятно, что дробление на отдельные сегменты зависит как от предполагаемого темпа речи, так и от нашей воображаемой выразительности. Мы приняли за правило членить на данном уровне текст с минимальной степенью дробления: различая вслед за Г.Н. Ивановой-Лукьяновой (1988: 10) «вариативные» и «обязательные» границы членения, мы отмечали лишь обязательные. Так, фразу S4t1f02 *Кот воровал у нас практически всё* мы посчитали односинтагменной, учитывая нераспространенность подлежащего. Реально же в спонтанной речи диктора было обнаружено на данном участке цепи две синтагмы — в отдельную синтагму выделено слово *кот*. Фраза *Я настолько себя плохо чувствую / я так устал / мне ничего не хочется* // при членении на «идеальные» синтагмы была поделена на три части. При **прослушивании** данной фразы также было выделено три синтагмы, с одной межсинтагменной паузой между второй и третьей синтагмами. Последнее деление мы отражаем на уровне Syntagma Real.

Как правило, смысловое единство в потоке речи соответствовало и интонационному единству вычленяемого отрезка звуковой цепи, хотя здесь возможны отклонения: нарушения интонационного единства за счет пауз хезитации, самоперебивов, ошибок и их самоисправлений (см. об этом, например, Степанова 2006). Кроме того, возможны индивидуальные отклонения в интонационном рисунке по сравнению с «классическими» ИК (Брызгунова 1980). Роль пауз при выделении (интонационном оформлении) синтагм в спонтанной речи — отдельная проблема, поэтому паузы вынесены в отдельный уровень.

Таким образом, уровень пауз (Pause) — «механический» уровень, единицами которого являются отрезки звукового потока от паузы до паузы. Паузой считались все перерывы в звучании, не являющиеся глухой смычкой согласных, а также заполненные хезитационные паузы.

При сопоставлении с уровнем Syntagma Ideal уровень Pause:

— служит для установления возможных/невозможных мест для пауз хезитации;

— помогает фиксировать случаи «сцепления» синтагм — случаи отсутствия физических пауз между «идеальными» синтагмами, т. е. смысловыми (и интонационными) единствами;

— необходим для фиксации и описания случаев, когда слова, входящие по смыслу в одну синтагму, образуют интонационное единство с другой (например, союзы).

5–6) Уровни слов (Word Real) и «идеальных» слов (Word Ideal).

Word Real — уровень фонетических слов (т. е. отрезков речевой цепи с одним словесным ударением, формируемый на основе уровня Sounds Real (см. ниже). Кроме того, на этом уровне учитываются все особенности фонетики спонтанной речи, в которой происходят различные модификации звуков в потоке речи (например, озвончение глухих, оглушение звонких согласных в нехарактерных для этого позициях, спирантизация смычных, оглушение сонантов, выпадение гласных и даже целых комплексов звуков[3]).

Word Ideal — орфографические слова, записанные в фонематической транскрипции.

Сопоставление уровней Word Real и Word Ideal позволяет описать систему энклитик спонтанной речи, а также ее морфонологические особенности.

7) Gramform — уровень морфологической разметки.

Словоформы, границы между которыми установлены на уровне word ideal, получают грамматическую метку — им приписывается морфологическое значение согласно морфологической модели, представленной в «Грамматическом словаре русского языка» А.А. Зализняка и грамматическом модуле программы starling (www.starling.rinet.ru), созданной С.А. Старостиным. (Для грамматической разметки использована вспомогательная база данных по нашему материалу, созданная С.А. Крыловым в программе Starling.)

8–9) Уровни морфем (Morpheme Real) и «идеальных» морфем (Morpheme Ideal). Morpheme Ideal — традиционное членение на морфемы (по словарю А. И. Кузнецовой и Т. Ф. Ефремовой) с поправкой на то, что мы имеем дело не с орфографической записью, а с фонематической транскрипцией. Morpheme Real — реализация морфем в речевом потоке, т. е. морфы.

Здесь мы сталкиваемся со следующими явлениями:

— конкретные морфы т. е. морфемы, имеющие «в идеале» различный «идеальный» фонетический состав,

[2] О некоторых проблемах, возникающих при выделении в спонтанных монологах синтагм, см.: Вольская, Степанова 2005: 16–24.

[3] См., например: Фонетика спонтанной речи 1988, а также раздел «Метапонятия языка и речи (к вопросу о поиске единиц описания живой речи)» (Асиновский и др.).

072-Ryko_A:_Layout 1 18.05.2008 16:13 Page 462

могут иметь одну и ту же фонетическую (акустическую) репрезентацию (например, одинаковое звучание морфем *-ыj* в словоформе *рыжий* и *-иjи* в слове *каждую*);

— проведение границ между морфами, аналогичных границам между соответствующими морфемами (что тоже может быть осложнено благодаря явлению фузии), порой оказывается просто невозможным.

Наряду с «классическими» случаями фузии (например, *казац-к-ий*: *казак-* + *-ск-* — фузия очевидна и на письме, *де[ц]-к-ий*: *дет-* + *-ск-* — фузия актуальна только для звучащей формы), в нашем материале «сплавление» может охватывать не только соседние звуки — оно может быть и более «глубоким». Так, в фразе S1t1f01 слово *обворовывал*, согласно традиционным представлениям, состоит из следующих морфем: *ab-var***-ov-yv-a**-*l-Ø*. Реально же звучит последовательность *ab-var-av:-l*, то есть справа от корня вместо набора различных суффиксальных морфем (*-ov-* аффикс, образующий глагольную основу, *-ыv-* суффикс итератива, *-a-* суффикс, образующий основу прошедшего времени/инфинитива) мы видим не поддающийся дальнейшему членению с формальной стороны звуковой комплекс *-av:-*.

— морфемы в некоторых случаях никак не реализуются (т.е. не вычленяется из звукового потока единица анализа морфемного уровня), грамматическое значение словоформы понимается из контекста. Это касается, прежде всего, окончаний прилагательных и местоимений. Например, во фразе *дети не знали как поймать рыжего кота / который у них посто… / э-э который постоянно у них что-то воровал продукты //* во фрагменте «*который у них»* окончание не выделяется, а предлог не реализован.

10) Уровень слогов (Syllabe).

В качестве рабочей гипотезы мы принимаем, что все слоги в русском языке открытые (закрытые возможны только в абсолютном конце перед паузой или в конце синтагмы).

11) Уровень ударения (Stress).

На этом уровне в единую шкалу (от 1 до 4) объединены словесное и синтагматическое ударение: слог, являющийся интонационным центром синтагмы или акцентно выделенного слова, отмечается максимальным баллом (4), ударный слог остальных фонетических слов синтагмы — баллом 3, первая степень редукции безударных — 2, вторая степень редукции — 1. Определение первой или второй степени редукции слога осуществлялось по общеизвестной формуле Потебни. Насколько эта формула отражает реально существующие темпоральные и динамические отношения между частями фонетического слова - сможет показать дальнейшее сравнение этого уровня с уровнем Sound Real.

12–13) Звуковые уровни представляют собой фонематическую (Sound Ideal) и фонетическую (Sound Real) транскрипции потока речи. Причем, если первая выполняется в соответствии с правилами, выработанными в рамках Санкт-Петербургской фонологической школы, учитывающей фонемную интерпретацию слов, произнесенных в полном типе, то вторая (фонетическая) транскрипция осуществляется на основе слухового анализа реального звучания отдельных звуков и/или звуков в контексте с помощью знаков Международного фонетического алфавита

— звуки (sound real);

— фонемы (sound ideal).

14) Уровень повторяющихся акустических элементов.

При анализе звучащего текста на русском языке мы опираемся на знания о фонетическом строе русского языка, о количественном и качественном составе фонетической системы и закономерностях ее функционирования в речи. Этими знаниями мы и пользуемся при транскрибировании звучащего текста. Например, признание звука [с] самостоятельной фонемой в русском языке заставляет нас транскрибировать акустическую последовательность *физическая пауза – высокочастотый шум* как [с] (см. рис. 1), а не как сочетание двух звуков [ts], как это будет, например, в английском языке.

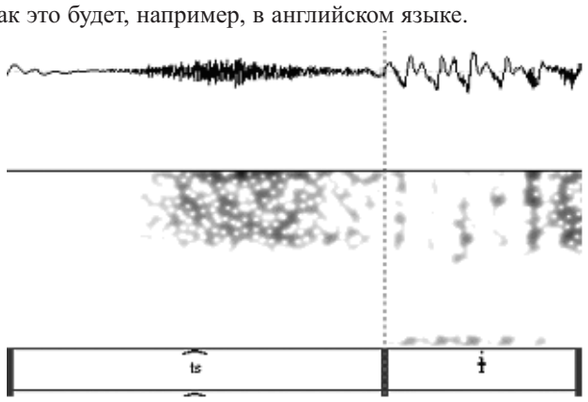


Рис 1. Осциллограмма, спектрограмма и транскрипция слога /са/ из слова нравится

072-Ryko_A:_Layout 1 18.05.2008 16:13 Page 463

Однако если мы анализируем записанный звучащий текст на неизвестном языке и нам недоступна информация, относящаяся к более высоким языковым уровням, то мы можем либо, опираясь на наш лингвистический опыт, давать участкам речевого потока приблизительную интерпретацию в рамках Международного фонетического алфавита, либо, обратившись к акустическим понятиям, членить речь на некие повторяющиеся акустические элементы (термин А.С. Асиновского) и давать им определенные метки.

Плодотворность второго подхода требует экспериментальной проверки.

Количество таких меток может быть различным: максимальным, учитывающим изменения звуковой волны по интенсивности, по длительности шумовой и/или гармонической составляющих, по месту и ширине формантных и шумовых полос и т. д.

В случае анализа нашего корпуса звучащих текстов мы пока ограничились минимальным составом вычленяемых акустических элементов, в равной мере пригодных для символьного транскрибирования звучащей речи как на известном, так и на неизвестном исследователю языке:

1) физическая пауза (глухая смычка) – (Ø);
2) высокочастотный шум при отсутствии периодов частоты основного тона (глухая щель) – (*fr)*;
3) наличие гармонических колебаний при отсутствии каких-либо других составляющих (звонкая смычка) (*Ø-v*);
4) гармонические колебания при наличии высокочастотного шума (звонкая щель) (*frv*);
5) гармонические колебания при наличии формантной структуры (вокалические участки) (*V).*

Разметка материала на этом уровне ориентирована на установление системных соответствий между результатами обработки речевого потока с помощью акустического речевого процессора и результатами экспертной разметки данного речевого потока.

На рис.2 и рис.3 представлен образец разметки корпуса на 14 описанных уровнях в программе Праат. Рис.2- разметка целой фразы *Я настолько себя плохо чувствую, я так устал, мне ничего не хочется*, рис.3 – разметка первой синтагмы той же фразы.

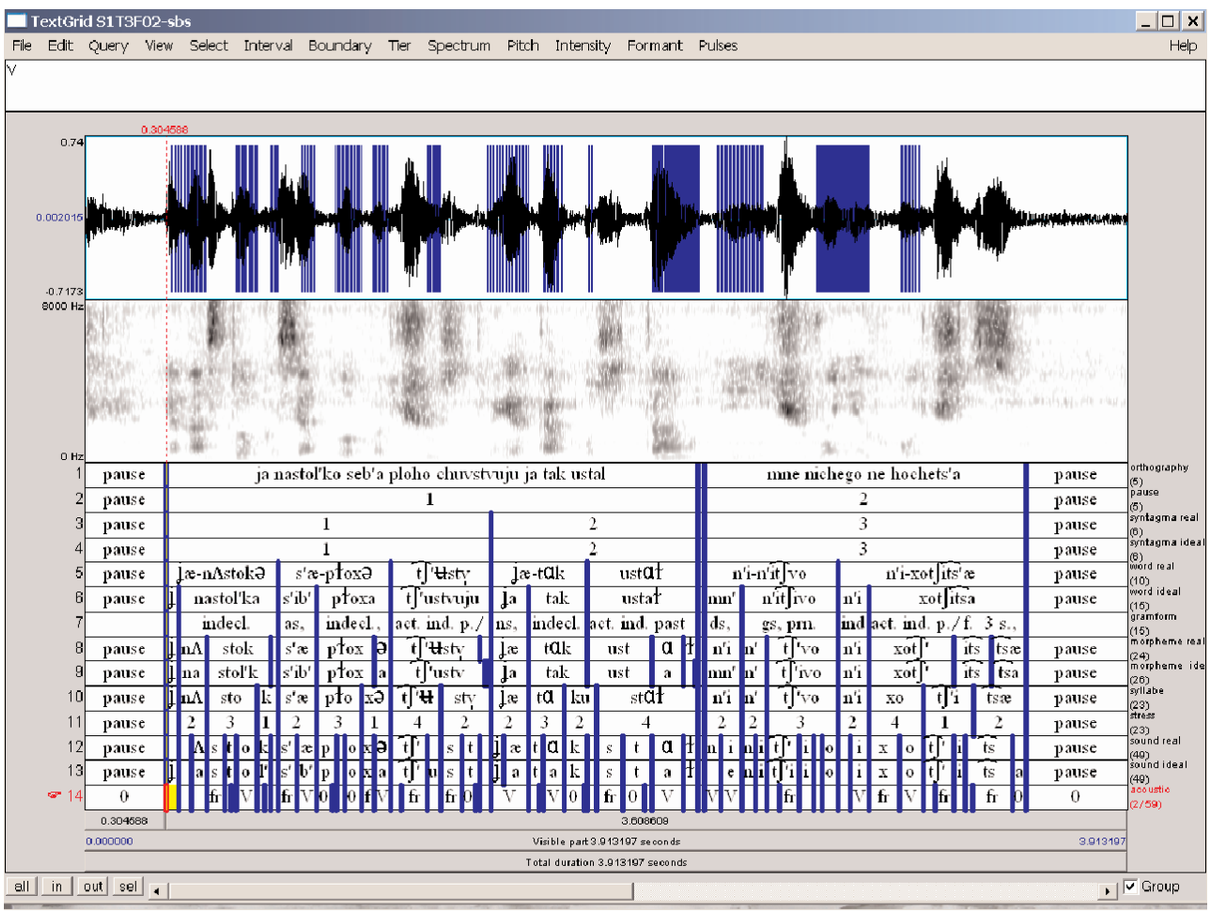


Рис. 2. Многоуровневая лингвистическая разметка фразы S1T3F02 (Я настолько себя плохо чувствую, я так устал, мне ничего не хочется)

072-Ryko_A:_Layout 1 18.05.2008 16:13 Page 464

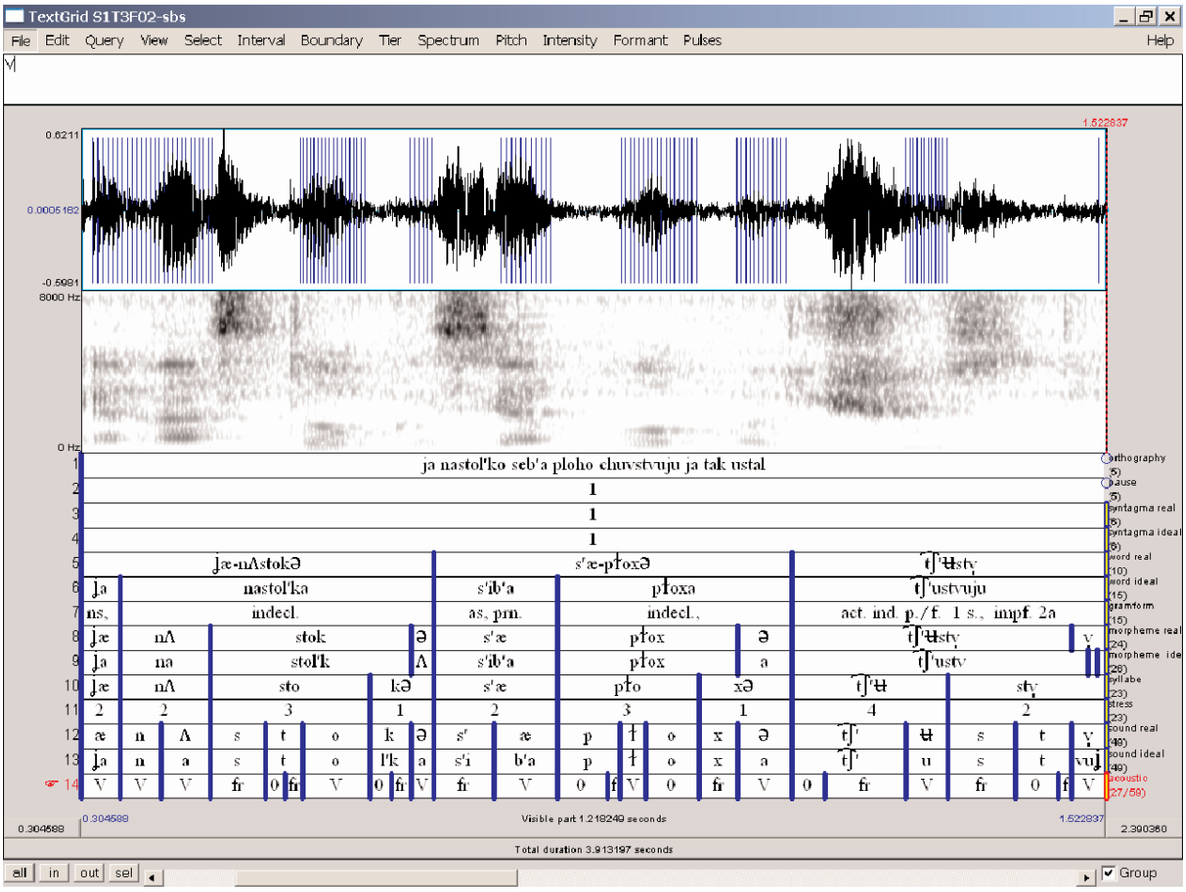


Рис. 3. Многоуровневая лингвистическая разметка той же фразы S1T3F02 (фрагмент)

Наличие в разметке уровней «реального» описания устных текстов на нескольких лингвистических уровнях и «идеального», ориентированного на представление того же текста в кодифицированном литературном языке, в полном типе произнесения, позволит в дальнейшем произвести с помощью подпрограмм Праата («скриптов») подробное сравнение этих уровней на материале представительного Звукового корпуса русского языка.

Список литературы

1. Асиновский А.С. и др. Полевая лингвистическая практика: Учебно-методический комплекс сложной структуры. Часть 2 // (В печати.)

2. Богданова Н.В. Метапонятия языка и речи (к вопросу о поиске единиц описания живой речи) // Материалы Международного симпозиума МАПРЯЛ «Инновации в исследованиях русского языка, литературы и культуры». Пловдив: 2006.

3. Бродт И.С. Спонтанный монолог в лингвистическом и социолингвистическом аспекта (на материале текстов разного типа): Автореф. дисс. ... канд. филол. наук // СПб.: 2007.

4. Брызгунова Е.А. Интонация // Русская грамматика. М.: 1980.

5. Вольская Н. Б., Степанова С. Б. О некоторых проблемах синтагматического членения спонтанного текста // XXX Международная филологическая конференция. Вып. 10. Секция фонетики. Ч. 1. СПб.: 2005.

5. Гордина М. В. Фонетика французского языка // Л.: 1973.

6. Зализняк А.А. Грамматический словарь русского языка // М.: 1977.

7. Иванова-Лукьянова Г.Н. Чтение вслух с опорой на пунктуацию // М., 1988

8. Кравченко М.Г., Зыкова М.А., Светозарова Н.Д., Братусь И.В. Ударение и интонация в немецком языке // Л., 1973.

9. Кузнецова А. И., Ефремова Т.Ф. Словарь морфем русского языка // М., 1986.

072-Ryko_A:_Layout 1 18.05.2008 16:13 Page 465

10. Степанова С.Б. Соотношение синтагматического и хезитационного членения в спонтанной русской речи // Материалы Международного симпозиума МАПРЯЛ «Инновации в исследованиях русского языка, литературы и культуры». Пловдив: 2006.

11. Фонетика спонтанной речи / под ред. Н. Д. Светозаровой // Л.:1988

12. Щерба Л.В. Фонетика французского языка // М.: 1955.

073-Savchuk_S:_Layout 1 18.05.2008 16:14 Page 466

ВАРИАНТНОСТЬ В РУССКОМ ЯЗЫКЕ. ПРОЕКТ СЛОВАРЯ
VARIATION IN RUSSIAN. DICTIONARY PROJECT

Савчук С.О. *(savsvetlana@mail.ru),* ***Гришина Е.А.*** *(rudi2007@yandex.ru)*
Институт русского языка им. В.В. Виноградова РАН

Статья представляет проект нового словаря вариантов в русском языке. Дается предварительное описание состава словника, типов задач, которые можно решать с помощью данного словаря, типы словарной информации, методы ее анализа.

Введение

В статье представлен проект нового словаря вариантов русского языка, который планируется создать на основе обследования состояния норм литературного языка конца XX – начала XXI века. Актуальность этой работы не вызывает сомнений, так как последнее массовое обследование состояния норм на основе анкетирования носителей русского языка проводилось в 60-е годы XX в. Его результаты нашли отражение в четырехтомной коллективной монографии «Русский язык и советское общество» (М., 1968) и книге «Русский язык по данным массового обследования» (М., 1974). В последующие десятилетия изучались отдельные фрагменты языковой системы, активные процессы в языке конца XX века[1], однако задача сплошного обследования всех «слабых участков» языковой нормы не ставилась. Все современные наблюдения имеют в значительной степени фрагментарный характер, что объясняется прежде всего ограниченностью материала, имевшегося в распоряжении исследователей.

Вариативность тесно связана с вопросами ортологии, поскольку существование нескольких вариантов обычно сопровождается их оценкой с точки зрения сложившихся норм употребления. Вариант может быть кодифицирован в литературном языке, оценен как более или менее предпочтительный, а может остаться за пределами литературной нормы. При этом отмечено закономерное несоответствие: сама вариантность может сохраняться неизменной в течение длительного времени, в то время как ее нормативные оценки меняются, причем достаточно быстро, иногда в течение десятилетий (Граудина 1980).

При этом следует учесть, что основные труды, кодифицирующие и регулирующие употребление вариантов, – академическая[2] и нормативно-стилистические грамматики[3], грамматический[4] и толковые словари, ортологические словари и справочники[5] – описывают состояние норм 40-50-летней давности[6]. Современные пособия, как правило, вторичны и компилятивны, в лучшем случае в них обновлены отдельные примеры.

Таким образом, и сегодня остается актуальной задача, сформулированная более четверти века назад: «Должна быть собрана, по возможности, полная коллекция всех основных типов грамматических вариантов и перечней форм-лексем – синтаксических, словоизменительных и словообразовательных, сопровождающаяся необходимыми нормативными оценками. Эта коллекция могла бы лечь в основу полной функционально-стилистической грамматики вариантов, которая была бы полезна как при подготовке новых изданий академической грамматики, так и при составлении ортологических словарей» (Граудина 1980: 272). Поэтому проектируемый словарь-справочник, подготовленный на новом материале, подробно описывающий новые явления и, возможно, дающий новые интерпретации и формулировки существующих правил, несомненно будет интересен и в научном, и в практическом плане.

Особенности жанра нового словаря-справочника

Новый словарь вариантов может быть отнесен к жанру «словарей трудностей» или «правильности русской речи». Однако в отличие от существующих пособий он будет иметь не столько нормативно-предписывающий,

[1] (Грамматические исследования 1989); (Грамматические исследования 1991); (Русский язык в его функционировании 1993); (Русский язык в его функционировании 1996); (Русский язык конца XX столетия 1996); (Русский язык сегодня 2006); (Современный русский язык 2008).
[2] (Русская грамматика 1980).
[3] (Розенталь1965).
[4] (Грамм 1977). В (Грамм 2003) словник словаря подвергся «умеренной модификации».
[5] (УРР); (ГПРР).
[6] Подробный обзор основных трудов, связанных с практической нормализацией грамматики см. в книге (Граудина 1980: 8-62)

073-Savchuk_S:_Layout 1 18.05.2008 16:14 Page 467

сколько объективно-регистрирующий характер и отвечать не на вопрос «Как правильно говорить по-русски?», а на вопрос «Как говорят по-русски». В качестве объективных показателей употребительности или распространенности варианта будут использованы частотные характеристики (ср.: Граудина 1980: 63).

Словник будет составлен с привлечением существующих ортологических словарей, наиболее полно отражающих вариативность языковых единиц и состав конкурирующих вариантов (см. список литературы). Состав включаемых в словарь единиц предполагается ограничить следующими типами вариантов:

- Варианты словоизменения, или морфологические варианты (*ко'рпусы – корпуса', жираф – жирафа, чая – чаю, мяукает – мяучит и под.*). Круг этих вариантов достаточно хорошо известен, но именно здесь происходят постоянные микроскопические изменения, не заметные невооруженным глазом, которые меняют границы этого круга и вызывают разноречивые рекомендации в нормативной литературе.
- Варианты словообразования (*инструктаж – инструктирование, желанный – желательный, колонизовать – колонизировать* и под.).
- Лексические варианты (бивак – бивуак, ветер – ветр, кирка – кирха, тамарисковый – тамариксовый и под.)[7].
- Синтаксические варианты: варианты предложного и беспредложного управления (*ехать на поезде – ехать в поезде – ехать поездом, скучать о ком – скучать по кому – скучать по ком, ждать автобус – ждать автобуса*).
- Акцентологические варианты слов и словоформ (*тво'рог – творо'г, булты'хнуться – бултыхну'ться, на'чался – начался'*).[8]

Поскольку русский литературный язык имеет развитую стилистическую систему, то анализ и оценка вариантов должны производиться с учетом их распределения по функциональным сферам, социальным, профессиональным разновидностям языка и т.д. В частности, варианты, распространенные в разговорной речи, могут быть нехарактерны для книжно-письменных текстов, и наоборот. Эта информация будет включена в состав словарной статьи в виде стилистических помет и/или комментариев.

Кроме того, в справочнике будет последовательно представлен диахронический аспект проблемы: отражены изменения в употреблении вариантов на протяжении XVIII–XX вв., которые выражаются в появлении или исчезновении вариантов, в смене соотношения между ними, в расширении или сужении сферы использования, изменении стилистических характеристик.

Наконец предполагается включить в словарную статью справочно-библиографическую информацию о характеристике слова или формы в основных нормативных словарях. Эта зона словарной статьи даст наглядное представление об изменении нормативных оценок вариантной формы на протяжении истории ее наблюдения.

Состав словника

Словник проектируемого словаря будет сформирован на основе основных лексикографических источников, описывающих с разной степенью подробности те или иные варианты. В состав словника будут включены варианты, которые упоминаются в большинстве существующих словарей и справочников, что может рассматриваться как свидетельство их распространенности и наличия проблемы в правилах их употребления. Кроме того, в словник войдут варианты, появившиеся в недавнее время и не получившие пока лексикографического описания и нормативной оценки, но зафиксированные в устных или письменных текстах. В настоящее время составлен предварительный словник, который будет в дальнейшем уточняться.

Классификация предварительного состава словника необходима с целью определения типов вариантных единиц и разработки типов словарных статей, соответствующих характеру описываемого явления. Некоторые варианты имеют индивидуальный характер (например, семантическое соотношение паронимов ноль–нуль), большая же часть вариантов (акцентологических, грамматических, паронимических, сочетаемостных) образует группы. Следовательно, описание индивидуального слова/группы слов в отдельной словарной статье должно сочетаться с описанием (вероятно, в отдельном разделе исследования) однотипных групп[9].

---

[7] Варианты второго и третьего типа составляют в большинстве своем паронимические пары или группы и традиционно рассматриваются в словарях трудностей, поскольку их семантическое и формальное сходство является причиной частых речевых ошибок, известных под названием «смешение паронимов».

[8] Эта задача может быть решена в ближайшее время по мере создания и пополнения акцентологического корпуса русского языка в составе НКРЯ.

[9] Предполагается, что описание синтаксических вариантов будет строиться в виде кратких очерков, описывающих модели синтаксических конструкций: согласования (подлежащего и сказуемого, определения и определяемого слова), обособленных деепричастных и адъективных оборотов и т.д.

073-Savchuk_S:_Layout 1 18.05.2008 16:14 Page 468

Реалистичность проекта

Очевидный вопрос, возникающий практически неизбежно, – располагаем ли мы ресурсами, способными предоставить нам объем материала, который был бы достаточным для получения достоверных данных, и средствами обработки материала, которые позволили бы решить задачу в обозримые сроки? Как представляется, современные электронные корпуса текстов, большие по объему и снабженные сложной лингвистической разметкой, как будто бы предназначены для изучения вопросов, связанных с соотношением узуса и нормы.

Материалом исследования служит прежде всего Национальный корпус русского языка (в отдельных случаях, при необходимости, будут привлекаться другие электронные ресурсы). Национальный корпус дает пользователю срез современного употребления русского языка, поэтому его в какой-то степени можно рассматривать как результат массового обследования, полученный однако не путем анкетирования, как это было в 1960-е годы, а путем целенаправленного отбора текстов[10]. То обстоятельство, что большинство изданных текстов, представленных в корпусе, подвергалось редактированию, не снижает их ценности для исследования, так как мы обследуем состояние **норм**, и редактор, корректирующий текст, выступает как носитель действующей нормы. Кроме того, в нашем распоряжении есть тексты, в которых редактирование либо отсутствовало (устная речь, интернет-коммуникация), либо было очень слабым (местная пресса, специальные журналы). Так что возможен сопоставительный анализ распространенности вариантов в разных типах изданий и источников.

Общий объем НКРЯ составляет в настоящее время около 160 млн словоупотреблений[11]. В нем можно условно выделить две части – современную и диахроническую. Современный корпус составляют тексты, период создания которых укладывается в рамки 1951-2007 гг. Объем этой части корпуса – 97,5 млн словоупотреблений, причем около половины текстов относится к периоду после 2000 г. Диахроническая часть составляет около 53 млн словоупотреблений и объединяет тексты XVIII в. (1,1 млн словоупотреблений), XIX в. (23,7 млн прозы и 2,5 млн в поэтическом корпусе) и 1-й пол. XX в. (25,7 млн).

Лингвистическая разметка (морфологическая и семантическая), которой снабжены тексты и словоформы, и поисковая система позволяют обследовать морфологические и таксономические классы слов, словообразовательные модели, синтаксические конструкции и отдельные словоформы и морфемы. Система метатекстовой разметки корпуса (автор, дата создания, сфера функционирования, жанр текста, библиографическое описание источника) позволяют определить координаты изучаемого лингвистического явления, отраженного в тексте, на функционально-стилистической и диахронической оси. Все эти возможности современных информационных ресурсов делают поставленную задачу вполне реальной и решаемой в обозримые сроки.

Приведем несколько примеров.

Изучение морфологических вариантов

Слово *корректив* во всех современных словарях (Грам, СОШ, Орф) относится к мужскому роду. Соответственно, форма Род. мн. этого существительного должна быть *коррективов*. В словаре ГПРР отмеченная в современных текстах форма Род. мн. *корректив* характеризуется как ненормативная со следующим комментарием: «Иногда нулевая флексия употребляется даже в тех случаях, в которых, казалось бы, она не должна появляться, например … «Простудное заболевание пока не внесло серьезных корректив в планы государственной деятельности президента» (РТР. «Вести». 1998. 13 мар.). Существительные *закоулок*, *корректив* зафиксированы в толковых словарях в мужском роде и должны принимать в род. мн. окончание *ов.* Однако эти существительные часто употребляются во множественном числе, поэтому исходная форма единственного числа им. падежа многими забывается и по аналогии с существительными, относящимися к потенциальным Pluralia tantum, у этих слов в род. мн. появляется нулевая флексия» (ГПРР, 180).

Данное объяснение могло бы нас удовлетворить, если бы речь шла о единичных случаях. Однако данные НКРЯ говорят о другом. Сплошной анализ контекстов употребления этого слова (всего 442 вхождения) выявил среди них такие, в которых представлены словоформы разной родовой принадлежности. Это формы Род. мн. (*коррективов* и *корректив*), Им.-Вин. ед. м. (*корректив*), Род. ед. м. (*корректива*), Вин. ед. ж. (*коррективу*),

[10] Как отмечала Л.К. Граудина, «метод пассивного наблюдения за устной и письменной речью можно считать идеальным с точки зрения безыскусственности условий собирания материала. Однако он не всегда может обеспечить репрезентативность выборки для вариантов с низкой частотой употребления» (Граудина 1980: 77). Если сравнить объем массива текстов, обследуемых в процессе создания словаря «Грамматическая правильность русской речи» (2 млн слов), с объемами современных корпусов, то замечание о репрезентативности выборки, по-видимому, должно быть пересмотрено.

[11] По данным сайта www.ruscorpora.ru на январь 2008 г.

073-Savchuk_S:_Layout 1 18.05.2008 16:14 Page 469

Тв. ед. м. и ж. (*коррективом* и *коррективой*). Количественное распределение форм мужского и женского рода по периодам показано в Таблице 2[12].

	XIX в. – 1-я пол. XX в. 49,4 млн	**2-я пол. XX в.** 97,5 млн
муж.	45	25
жен.	0	24

Таблица 2.

Таким образом, почти равное соотношение форм мужского и женского рода в текстах 2-ой половины XX в. свидетельствует о появлении у слова **корректив,** -а, *м.* варианта **корректива,** -ы, *ж.,* и эта пара пополняет группу существительных, имеющих колебание в роде (*жираф – жирафа, компонент – компонента* и пр.). Хотя объяснение причин возникновения вариантов требует специального исследования, выскажем все-таки осторожное предположение, что причина в данном случае может состоять в том, что во 2-ой половине XX в. это слово, резко повысив свой частотный ранг (из всех 442 вхождений слова в корпус 377 случаев, или 85%, приходится на тексты 1990-х гг., при этом объем корпуса современных текстов всего в 2 раза больше), употребляется в конструкциях с формами множественного числа, в которых снято противопоставление по роду (самая массовая форма – Им.-Вин. мн. – 312 случаев). Поэтому по аналогии с такими словами, как *инициатива, директива, инвектива* и нек. др., относящимися к книжному слою лексики и встречающимися в сходных контекстах, форма *коррективы* начинает осмысляться как Им.-Вин. мн. существительного женского рода *корректива*. Однако это предположение нуждается в проверке на материале всей группы существительных с колебанием в роде, что, кстати, будет способствовать построению прогнозов относительно судьбы конкурирующих вариантов.

Изучение словообразовательных вариантов и паронимов

Следующий пример демонстрирует возможности корпуса при описании соотношений между однокоренными словами в составе группы паронимов *жалобный, жалостный, жалостливый, жалкий*. Различия, по данным толковых словарей, сводятся к следующему:

Возбуждающий, внушающий жалость	Испытывающий жалость, склонный к жалости
Жалобный. Выражающий жалобу, скорбь, тоску. *Ж. писк. Жалобно просить*. **Жалостный** (разг.). 2) То же, что жалобный. *Ж. голос*. **Жалкий.** 1) Возбуждающий жалость, несчастный, беспомощный. *Ж. вид. Ж. фигура*. *Жалко улыбнуться*. 2) Жалобный, трогательный. *Говорить ж. слова*	**Жалостливый** (разг.). Склонный к жалости, сострадательный. *Ж. взгляд*. **Жалостный** (разг.). 1) Жалостливый, соболезнующий. *Ж. взгляд*.

Количественный анализ употребления этих слов по различным хронологическим срезам корпуса выявил следующую картину[13]. По каждому периоду в левой графе представлено общее количество употреблений, в правой – частота на 1 млн словоупотреблений.

[12] Здесь, как и в других примерах, важно не точное значение количественных показателей, т.к. оно может меняться по мере пополнения корпуса новыми текстами. Принципиальное значение имеет соотношение между количественными параметрами, позволяющее проследить тенденцию и динамику развития изучаемых явлений.

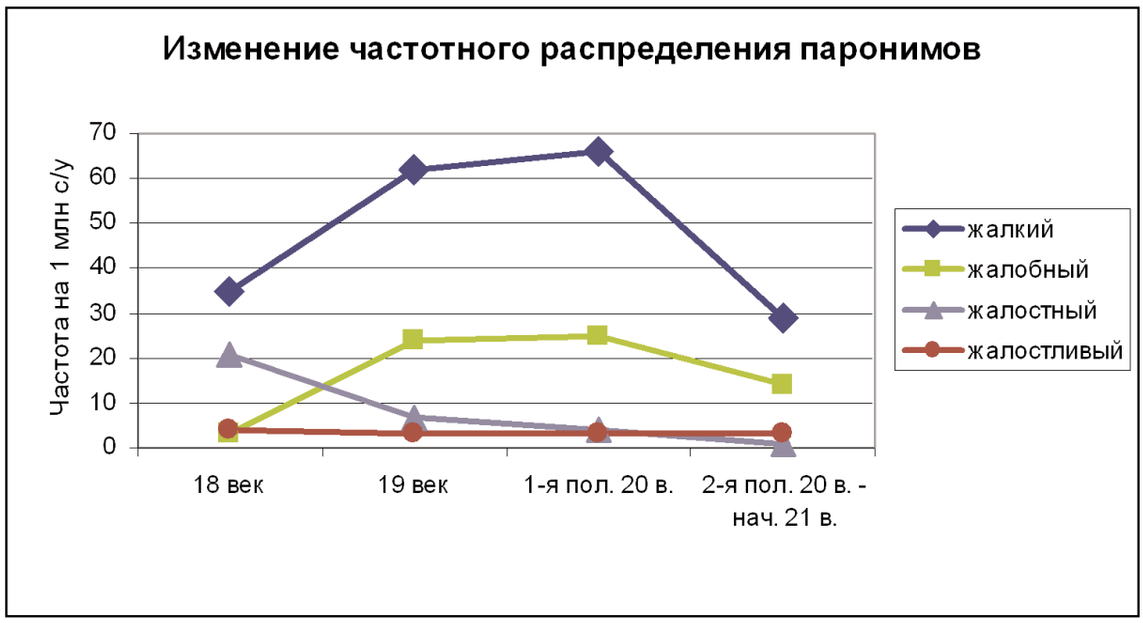
[13] Рассматривались контексты, в которых представлены не только формы прилагательных, но и наречия, которые составляют примерно половину всех употреблений. Однако из рассмотрения были исключены контексты с предикативом *жалко*, составляющие от половины до двух третей (в текстах 2001-2007 гг.) всех употреблений слова *жалкий*.

073-Savchuk_S:_Layout 1 18.05.2008 16:14 Page 470

		1700-1799 1,1 млн		**1800-1899** 23,7 млн		**1900-1949** 25,7 млн		**1950-2000** 49,8 млн		**2001-2007** 46,8 млн	
Жалобный		3	3	571	24	641	25	1002	21	338	7
Жалкий		35	35	1474	62	1690	66	2003	41	822	18
Жалостный	‘внушающий жалость’	18	21	169	7	104	4	97	2	19	0,4
	‘испытывающий жалость’	3		4		5		6		-	
Жалостливый	‘испытывающий жалость’	4	4	75	3	76	3	185	4	75	2
	‘внушающий жалость’	-		7		5		25		19	

Таблица 3.

Изменение частотности рассматриваемых прилагательных в диахронической перспективе представлено на графике.



Жалобный и *жалкий* показали стабильно высокую частоту употребления на протяжении всего периода, причем частотность прилагательного *жалкий* значительно выше, что объясняется его более сложной семантической структурой. Напротив, прилагательное *жалостный*, широко представленное в текстах XIX в., обнаруживает тенденцию к неуклонному снижению частоты, вплоть до значений меньших единицы на миллион словоупотреблений в современных текстах. О чем это может свидетельствовать?

Содержательный анализ контекстов употребления слова *жалостный* показывает, что из двух словарных значений оно преимущественно реализует одно, а именно значение ‘внушающий жалость’: *жалостный вопль, писк, стон; жалостная песня, мелодия; жалостное зрелище, состояние.* Но здесь ему сильную конкуренцию составляют прилагательные *жалобный* и *жалкий* (последнее в процессе употребления приобретает в ряде контекстов сильный оценочный компонент), и график показывает, кто побеждает в этой конкурентной борьбе. Что касается второго значения прилагательного *жалостный,* ‘испытывающий жалость’, то оно реализуется в незначительном количестве контекстов (в текстах 2001-2007 гг. не отмечено вовсе), и здесь преимущество у прилагательного *жалостливый*.

073-Savchuk_S:_Layout 1 18.05.2008 16:14 Page 471

Это прилагательное претерпело, пожалуй, самые значительные трансформации. Вопреки тому, что в словарях (Ушаков, СОШ) оно зафиксировано в единственном значении ‘склонный к жалости’, уже в текстах XIX в. его можно встретить в значении ‘внушающий жалость’, то есть в одном из значений его ближайшего соседа по паронимическому ряду – прилагательного *жалостный*[14].

По-видимому, это уже свидетельствует не о распространенной речевой ошибке (смешение паронимов), а о формировании у прилагательного *жалостливый* второго значения и об уподоблении семантических структур двух паронимов. Этот факт уже отмечен в (Розенталь, Теленкова 2007), где слова *жалостный* и *жалостливый* приводятся в составе паронимического ряда, оба с пометой *разг.* и с одинаковым набором значений.

Изучение вариантов глагольно-именного управления

Вариантность в этой зоне можно показать на примере группы глаголов *беспокоиться, тревожиться, волноваться, переживать,* которые относятся к одному синонимическому ряду с общим значением ‘испытывать неприятное чувство, какое обычно бывает, когда человеку неизвестно что-то важное о ситуации, которая его касается, и когда он опасается, что ситуация изменилась или может измениться к худшему’[15].

Все рассматриваемые глаголы допускают варианты в управлении зависимым существительным, и именно эта вариативность получает в словарях нормативную оценку. Приведем эти описания по данным (Син) и (УРР).

Наиболее полно варианты управления описаны в словаре синонимов (порядок следования моделей соответствует степени предпочтительности варианта).

Словарь	**Глагол**	**Без дополнения**	**За + Вин. п.**	**О/Об/Обо + Предл. п.**	**Из-за, насчет, по поводу + Род. п.**	**Придаточное предложение**
Син	**беспокоиться**	+	кого	о ком	чего	+
	тревожиться	+	кого	о ком	чего	+
	волноваться	+	кого	–	чего	+
УРР	**беспокоиться**	нет данных	кого /что *разг.*	о ком/чем	нет данных	нет данных
	тревожиться	нет данных	кого /что	о ком/чем *разг.*	нет данных	нет данных
	волноваться	нет данных	кого /что	о ком/чем	нет данных	нет данных
	переживать	нет данных	кого /что *прост.*	о ком/чем *разг.*	кого /чего	нет данных

Таблица 4.

Как видно из таблицы, рекомендации словарей отличаются в оценке вариантов управления. Так, в (Син) указываются ограничения на семантику управляемого существительного (напр., **беспокоиться за** + кого: объект беспокойства – живое существо; **беспокоиться из-за** + кого: причина беспокойства – обстоятельства, события). В (УРР) по сравнению с (Син) более строгие стилистические предписания при более свободных требованиях к семантике. Отвечают ли эти предписания данным, полученным с помощью Национального корпуса русского языка?

Детальный качественный анализ глагольного управления в рассматриваемой группе проведен на основе ручной обработки контекстов употребления этих глаголов в составе нескольких подкорпусов:

- 1800-1899 XIX-художественные тексты; XIX-нехудожественные тексты

[14] Ср.: Лизанька, – сказал Манилов с несколько жалостливым видом, – Павел Иванович оставляет нас! – Потому что мы надоели Павлу Ивановичу, – отвечала Манилова. [Н.В. Гоголь. Мертвые души (1842)]. Но тут великий комбинатор принялся молоть такую жалостливую чепуху, что работник связи ушел в другой зал искать посылку бедного студента. [Илья Ильф, Евгений Петров. Золотой теленок (1931)]. Во 2-ой половине XX в. количество этих употреблений резко возрастает, расширяется круг сочетаемости прилагательного жалостливый. Ср.: С этими словами Панюшкин вытащил из шкафа жалостливого, лысого и во многих местах штопанного игрушечного медведя. [Валерий Панюшкин. За Виню // «Столица», 1997.02.17]. Она то покрикивает на меня, то жалостливо просит пощадить ее, уставшую от ожидания и опоздавшую на обед. [Рубин Евгений. Пан или пропал. Жизнеописание (1999-2000)].

[15] См. (Син.).

073-Savchuk_S:_Layout 1 18.05.2008 16:14 Page 472

- 1900-1949 XX-1-художественные тексты; XX-1-нехудожественные тексты
- 1950-2005 XX-2-художественные тексты; XX-2-устные тексты; XX-2-тексты электронной коммуникации.

Количественные результаты приведены в (Савчук 2007), здесь остановимся на общих выводах. Анализ соотношения глаголов рассматриваемого синонимического ряда по данным НКРЯ показывает, что доминантой ряда, по-видимому, справедливо считается глагол *беспокоиться* как наиболее универсальный, и это лидирующее положение в группе он занимает на протяжении всего периода.

Глагол *тревожиться*, напротив, самый малочастотный, с тенденцией к снижению употребительности (в устных текстах и в электронной коммуникации он не встретился ни разу). Современным носителем языка он ощущается как книжный, литературный.

Глагол *переживать*, не отмеченный в составе рассматриваемого ряда в (Син), закрепился в языке в значении ‘волноваться, беспокоиться о ком-чем-н.’ во 2-ой половине XX века. При этом он изменил свои грамматические свойства, перейдя в разряд непереходных. Первые единичные случаи такого употребления отмечаются в художественных текстах 1-ой половины XX века (например, Ильф и Петров, «Золотой теленок», 1927), для текстов 1960-1980-х годов это уже массовое явление, отмеченное, например в (Шмелев 1962, 442).

В современной разговорной речи (в устных текстах и электронной коммуникации) употребительность этого глагола резко превышает употребление всех остальных глаголов рассматриваемой группы. В целом глагол *переживать* сохраняет оттенок разговорности, хотя сферы его употребления стремительно расширяются.

Анализ вариантов глагольного управления показывает, что все глаголы обнаруживают тенденцию к выравниванию моделей управления.

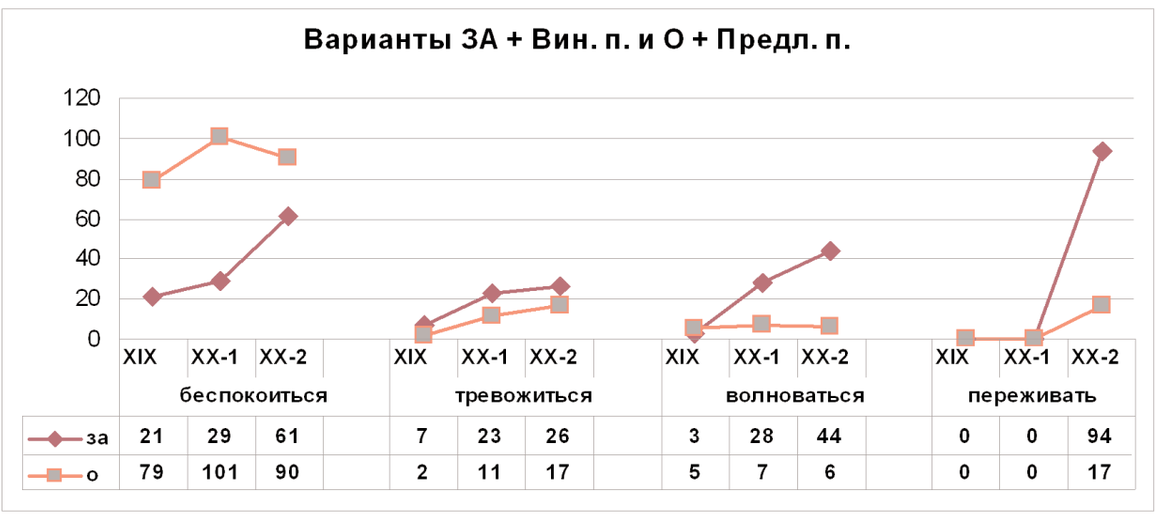
Для глагола *беспокоиться* наблюдается рост варианта управления с предлогом ЗА+Вин. п. Для текстов XX века этот вариант практически равноправен с вариантом О+Предл. п. Следует отметить также значительный рост в этот период вариантов с предлогом НАСЧЕТ + Род. п.

Для глагола *тревожиться* наблюдается тенденция к сокращению дистанции между вариантами *тревожиться за кого-что/ тревожиться о ком-чем*.

Для глагола *волноваться* вариант *волноваться за кого-что* предпочтительнее, чем *волноваться о ком-чем,* причем на протяжении XX века этот разрыв увеличивается. Однако не следует вообще игнорировать этой модели, как это делается в (Син), поскольку она стабильно представлена в корпусе на протяжении всего рассматриваемого периода, в том числе и в современной устной речи и электронной коммуникации. Следует также отметить активность конструкции *волноваться из-за кого* и чаще *чего* (указание на причину).

Для глагола *переживать* предпочтительнее вариант *переживать за кого-что, переживать из-за кого-чего* и реже встречается *переживать о ком-чем,* хотя без дополнения этот глагол употребляется чаще.

В целом можно сделать вывод, что вариант управления V+ЗА+Вин. п. оказывается более продуктивным, чем другие: на протяжении XIX-XX вв. происходит количественное сближение вариантов V+ЗА+Вин. п. и V+О+Предл. п. у тех глаголов, у которых в начале периода преобладал вариант V+О+Предл. п. (*беспокоиться, тревожиться*) и увеличение количественного разрыва у глагола с изначальным преобладанием варианта V+ЗА+Вин. п. (*волноваться*).



073-Savchuk_S:_Layout 1 18.05.2008 16:14 Page 473

Сравнивая полученные результаты с рекомендациями словарей и справочников, приходим к заключению, что эти рекомендации выглядят излишне жесткими, не вполне отвечают реальному употреблению, а потому часто не выполняются. В частности, помета *прост*., приписанная модели *переживать за кого-что* в (УРР) и тем самым запрещающая использовать данную конструкцию в литературной речи, не соответствует реальности: модель широко используется в художественной литературе (Б. Васильев, В. Тендряков, В. Гроссман, В. Чивилихин, И. Грекова, В. Шукшин, А Рыбаков и др.), причем не только в речи персонажей, но и в авторской речи.

Кроме того, результаты анализа контекстов показали, что семантическая сочетаемость глаголов с существительными в составе конструкций более свободная, чем это описано в (Син), шире круг предлогов и союзных средств, позволяющий создавать с этими глаголами различные синтаксические конструкции.

***

Таким образом, корпусной подход к исследованию вариантов различных типов позволяет провести количественный и качественный анализ вариантов, выявить тенденции в соотношении вариантов, проследить появление новых явлений, их развитие и перспективы, уточнить существующие описания и внести коррективы в рекомендации нормативных словарей и справочников.

Список литературы

1. ГПРР – Граудина Л.К., Ицкович В.А., Катлинская Л.П. Грамматическая правильность русской речи. М.: Наука, 1976; 3-е изд. 2004

2. Грамм 1977 – Зализняк А.А. Грамматический словарь русского языка. Словоизменение. М.: Русский язык, 1977

3. Грамм 2003 – Зализняк А.А. Грамматический словарь русского языка. Изд. 4-е, испр. и доп. М.: 2003

4. Грамматические исследования 1989 – Грамматические исследования: функционально-стилистический аспект. / Отв. ред. Д.Н. Шмелев. М., 1989

5. Грамматические исследования 1991 – Грамматические исследования: функционально-стилистический аспект. / Отв. ред. Д.Н. Шмелев. М., 1991

6. Граудина 1971 – Граудина Л.К., Ицкович В.А., Катлинская Л.П. Грамматические варианты: Опыт частотного словаря. М.: Наука, 1971

7. Граудина 1980 – Граудина Л.К. Вопросы нормализации русского языка: Грамматика и варианты. М.: Наука, 1980.

8. Еськова – Еськова Н.А. Краткий словарь трудностей русского языка. Грамматические формы. Ударение. М.: Русский язык, 2003

9. Орф – Русский орфографический словарь. М., 2005

10. Розенталь 1965 – Розенталь Д.Э. Практическая стилистика русского языка. М., 1965 //Розенталь Д.Э. Русский язык: Справочник-практикум. М.: Оникс; Мир и Образование, 2007

11. Розенталь, Теленкова 2007 – Розенталь Д.Э., Теленкова М.А. Словарь трудностей русского языка. М.: Айрис-пресс, 2007

12. Русская грамматика 1980 – Русская грамматика /Под ред. Н.Ю. Шведовой и др. Т. 1-2. М.: Наука, 1980

13. Русский язык в его функционировании 1993 – Русский язык в его функционировании: коммуникативно-прагматический аспект. / Отв. ред. Е.А. Земская и Д.Н. Шмелев. М., 1993

14. Русский язык в его функционировании 1996 – Русский язык в его функционировании: уровни языка. / Отв. ред. Д.Н. Шмелев и М.Я. Гловинская. М., 1996

15. Русский язык и советское общество: Морфология и синтаксис современного русского литературного языка. М.: Наука, 1968

16. Русский язык конца XX столетия 1996 – Русский язык конца XX столетия (1985—1995) / Отв. ред. Е.А. Земская. М.: Языки славянской культуры, 1996

17. Русский язык по данным массового обследования. М.: Наука, 1974

18. Русский язык сегодня 2006 – Русский язык сегодня 4. Проблемы языковой нормы: Сб. статей. М.: Азбуковник, 2006

19. Савчук 2007 – Savchuk Svetlana. Corpus-based Investigation of Language Change: the Case of RNC // Proceedings of the Corpus Linguistics Conference CL2007 University of Birmingham, UK, 27-30 July 2007 /Matthew Davies, Paul Rayson, Susan Hunston, Pernilla Danielsson (eds.) http://ucrel.lancs.ac.uk/publications/CL2007/-final/181/181_Paper.pdf

073-Savchuk_S:_Layout 1 18.05.2008 16:14 Page 474

20. Син – Новый объяснительный словарь синонимов русского языка / Под общим руководством Ю.Д. Апресяна. Москва-Вена, 2004

21. Современный русский язык 2008 – Современный русский язык: Активные процессы на рубеже XX-XXI веков / Отв. ред. Л.П. Крысин. М.: Языки славянской культуры, 2008

22. СОШ – Ожегов С.И., Шведова Н.Ю. Словарь русского языка. М.: 1999

23. Трудности словоупотребления и варианты норм русского литературного языка: Словарь-справочник / Под ред. К.С. Горбачевича. Л.: Наука, 1974

24. УРР – Розенталь Д.Э. Управление в русском языке. М.: Русский язык, 1981 // Розенталь Д.Э. Русский язык: Справочник-практикум. М.: Оникс; Мир и Образование, 2007

25. Ушаков – Толковый словарь русского языка. / Под ред. Д.Н. Ушакова. М.: Русские словари, 1994

26. Чернышев 1911 – Чернышев В. Правильность и чистота русской речи: Опыт русской стилистической грамматики. СПб, 1911

27. Шмелев 1962 – Шмелев Д.Н. Некоторые вопросы развития и нормализации современного русского языка. 1962 // Шмелев Д.Н. Избранные труды по русскому языку. М.: Языки славянской культуры, 2002.

.

074-SidorovaE:_Layout 1 18.05.2008 16:15 Page 475

МНОГОЦЕЛЕВАЯ СЛОВАРНАЯ ПОДСИСТЕМА ИЗВЛЕЧЕНИЯ ПРЕДМЕТНОЙ ЛЕКСИКИ

MULTIPURPOSE DICTIONARY SUBSYSTEM FOR EXTRACTION OF SUBJECT LEXICON

Сидорова Е.А. *(lena@iis.nsk.su)*
Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск

Рассматривается словарная технология, предназначенная для создания предметно-ориентированных словарей и их использования при решении различных задач анализа текста в информационных системах. Исследуются проблемы одновременного использования нескольких словарей и согласования разнотипной словарной информации.

1. Введение

Современные требования к качеству содержательного анализа текстовых документов приводят разработчиков к необходимости ограничивать как предметную область содержания документа, так и стиль оформления текста. Кроме того, большинство информационных систем, если они не являются специализированным литературным ресурсом, работают с документами, написанными в жанре «деловой прозы» [1]. Однозначный контекст, строгое выражение мысли в таких текстах значительно уменьшает количество способов выражение одной и той же информации. Служебная либо общезначимая лексика, характерная для текста на естественном языке, исполняет роль связывания или замещения значимой лексики; не используются образные, литературные, эмоциональные выражения. Безусловно, это упрощает создание словарей – позволяет создавать специализированные словари небольшого объема и эффективно их использовать.

С другой стороны, разнообразие целей (задач) анализа текста даже в рамках одной информационной системы требует включать в словари самую разнообразную информацию. Например, при создании информационного Интернет-портала [2] для решения различных задач потребовалась следующая информация:

- для оценки релевантности документа тематике портала – статистика, многословные устойчивые термины, сокращения и аббревиатуры предметной области и т.п.;
- для автоматического определения жанра документа – признаки сегментов, лексические конструкции, формальные модели документов;
- для классификации или рубрикации – поддержка иерархии признаков и статистика;
- для простой индексации (словами) – морфология;
- для содержательной индексация – семантические признаки, связи между терминами (синонимы, антонимы и т.п.);
- для обработки незнакомых слов (поиск и исправление ошибок) – хэш-словари словоформ.

На практике такой словарь зачастую представляет собой объединение нескольких разнотипных и разноформатных словарей, к тому же создаваемых разными авторами или группами авторов. Использование набора словарей для реализации конкретной задачи ставит вопрос о согласовании данных из разных словарей (таких как семантические характеристики, связи между элементами словарей и т.п.), а также согласовании получаемых результатов словарной обработки текста с помощью словарных компонентов.

В данной работе описывается система, назначение которой создание ПО-словарей и их использование для анализа текста в различных информационных системах.

2. Информационное наполнение словарей

Одной из задач разрабатываемой технологии было создание гибких механизмов, позволяющих специалисту проводить тонкую настройку структуры словаря. Это выражается в возможности в значительной степени формировать структуру словарной статьи для терминов одного вида.

074-SidorovaE:_Layout 1 18.05.2008 16:15 Page 476

2.1. Типизация признаков словарной статьи

Все признаки, хранящиеся в словарной статье термина, мы условно разделили на четыре группы, в зависимости от их функционального назначения.

Терминообразующие признаки служат для того, чтобы с одной стороны выявить термин в тексте (анализ), с другой – послужить основой для построения терминов (синтез). Для разных видов терминов, набор терминообразующих признаков различен. В случае терминообразующих признаков возможность изменять структуру словарной статьи сильно зависит от вида термина.

Семантические признаки приписываются терминам словаря и передаются внешним программам вместе с найденными в тексте терминами. Эта семантическая информация позволит связать элементы словаря с онтологическими классами проблемной и предметной области и в дальнейшем будет использоваться на стадии семантического анализа текста. Набор признаков и их тип определяется пользователем при создании и наполнении словаря.

Семантическая информация может быть выражена в словаре по-разному. Во-первых, это семантический класс (или несколько классов), к которому приписывается термин словаря. Во-вторых, это атрибуты, добавляемые пользователем в словарную статью терминов. Наличие и тип значения атрибута фиксируется в структуре словарной статьи для всех или выделенной группы терминов (класса) словаря. В-третьих, это связи между элементами словаря, такие как отношения синонимии, омонимии, часть-целое и пр.

Статистические признаки накапливают статистическую информацию о появлении термина в обрабатываемых текстах. Часто такие признаки служат для наполнения словарей. Так, при создании информационных систем, как правило, изначально имеется большая выборка ресурсов, размеченная и соотнесенная тематическим разделам, и, используя классические методы обучения, можно сразу получить начальное наполнение словаря, которое, в противном случае, пришлось бы вводить вручную многочисленным специалистам. Помимо этого, наличие в словаре такой информации позволяет использовать статистические методы классификации (рубрикации) для определения общей тематики документа.

Если в случае семантических признаков, пользователь имеет полную свободу в формировании словарной статьи, то в случае статистических терминов, такая свобода отсутствует. Пользователь может только задавать иерархию рубрик или тем, согласованных с обучающей выборкой, в соответствии с которой будет автоматически накапливаться статистика, и, при желании, вручную устанавливать веса терминов в той или и иной теме, которые будут при дальнейшей обработке иметь приоритетное значение по сравнению с автоматически получаемыми значениями.

Еще одна важная группа признаков, упрощающая работу со словарями внешним программам, это динамические признаки, которые появляются у терминов после того, как они найдены в тексте документа в результате словарного анализа текста. К ним относятся такие признаки, как статус обновления статистических параметров термина, видимость найденных терминов после словарной обработки текста (см. Рис.1.), позиция найденного термина в тексте и др. К этой же группе можно отнести и те признаки, которые влияют на значения динамических признаков. Например, значимость термина позволяет управлять параметром видимости. Этот атрибут имеет следующие значения:

- самостоятельно незначимый термин всегда «невидим» и используется только в составе других терминов. Например, аббревиатуры в составе других названий (ОАО, СО);

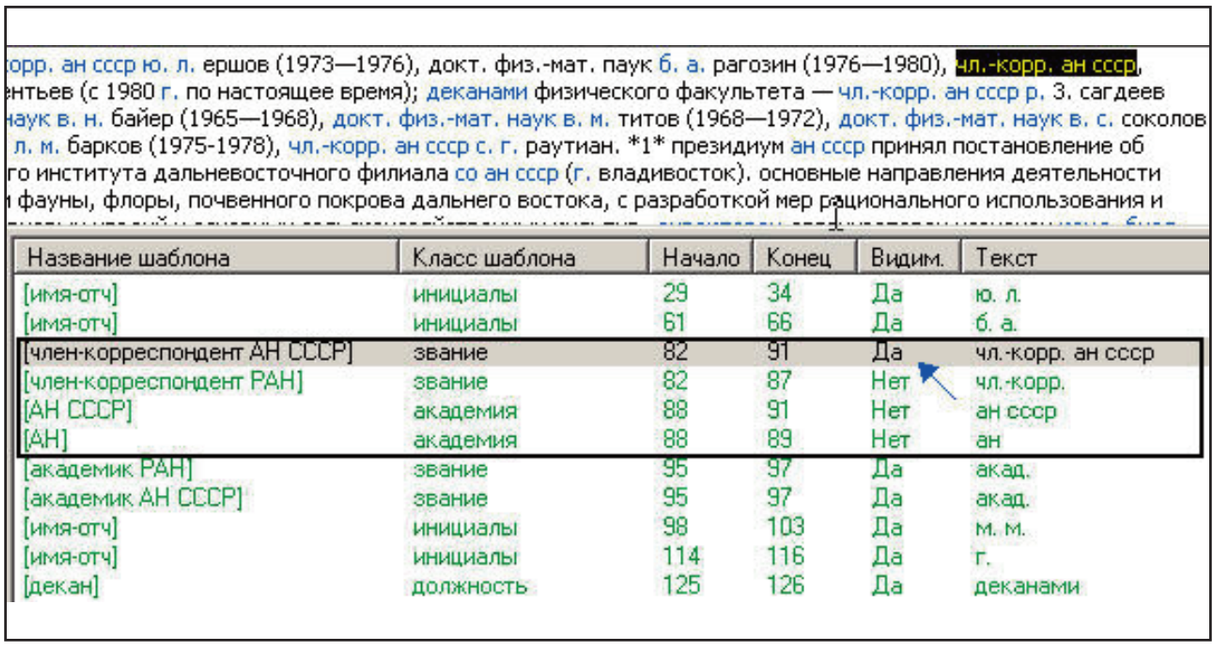
- самостоятельно значимый термин имеет положительную видимость, если не входит в состав другого самостоятельно значимого или абсолютно значимого термина и отрицательную видимость в противном случае (это означает, что строка текста, покрываемая термином, строго вложена в часть текста, покрываемом другим термином). К таким терминам относится большинство предметных терминов;

- универсально значимый термин является самостоятельно значимым, но не влияет на видимость вложенных в него терминов (например, разрывные термины или термины, характеризующие целый текстовый сегмент);

- абсолютно значимый термин всегда имеет положительную видимость независимо от того, входит он в состав другого термина или нет (например, такие жанровые термины как обращение, заключительная реплика и т.п.).

На Рис.1. приведен пример результата словарной обработки текста.

074-SidorovaE:_Layout 1 18.05.2008 16:15 Page 477



орр. ан ссср ю. л. ершов (1973—1976), докт. физ.-мат. паук б. а. рагозин (1976—1980), чл.-корр. ан ссср,
нтьев (с 1980 г. по настоящее время); деканами физического факультета — чл.-корр. ан ссср р. з. сагдеев
наук в. н. байер (1965—1968), докт. физ.-мат. наук в. м. титов (1968—1972), докт. физ.-мат. наук в. с. соколов
л. м. барков (1975-1978), чл.-корр. ан ссср с. г. раутиан. *1* президиум ан ссср принял постановление об
го института дальневосточного филиала со ан ссср (г. владивосток). основные направления деятельности
фауны, флоры, почвенного покрова дальнего востока, с разработкой мер рационального использования и

Название шаблона	Класс шаблона	Начало	Конец	Видим.	Текст
[имя-отч]	инициалы	29	34	Да	ю. л.
[имя-отч]	инициалы	61	66	Да	б. а.
[член-корреспондент АН СССР]	звание	82	91	Да	чл.-корр. ан ссср
[член-корреспондент РАН]	звание	82	87	Нет	чл.-корр.
[АН СССР]	академия	88	91	Нет	ан ссср
[АН]	академия	88	89	Нет	ан
[академик РАН]	звание	95	97	Да	акад.
[академик АН СССР]	звание	95	97	Да	акад.
[имя-отч]	инициалы	98	103	Да	м. м.
[имя-отч]	инициалы	114	116	Да	г.
[декан]	должность	125	126	Да	деканами

Рис. 1. Пример разбора цепочки вложенных терминов.

Найденные термины представляются в виде объектов, у которых, в данном случае, имеются динамические признаки *Начало, Конец, Видимость* и *Текст*. Границы терминов *АН, АН СССР, член-корреспондент РАН* (элемент *РАН* необязательный) включены в интервал, определяемый границами термина *член-корреспондент АН СССР*, поэтому для этих терминов признак видимости отрицателен. Отрицательное значение признака *видимость* означает, что термин не будет принимать участие в дальнейшем анализе.

2.2. Типизация терминов

Нами были выделены несколько видов терминов, для каждого из которых реализован отдельный словарный компонент.

Лексема или однословный термин представляет собой слово во всей совокупности его форм и значений. Словарная статья лексемы содержит следующие терминообразующие признаки:

- нормальная форма,
- основа – неизменяемая при склонении или спряжении часть лексемы (у некоторых лексем основа может быть пустой, тогда лексема определяется парадигмой),
- парадигма – набор псевдофлексий или изменяемых частей слова,
- морфологический класс (класс, в частности, включает часть речи и словообразующие морфологические признаки лексемы),
- тип: слово универсального словаря, предсказание, слово служебного словаря и т.п.

Набор морфологических признаков, включаемых в словарную статью лексемы, может легко настраиваться пользователем, являющимся специалистом в данной области.

Термин-словокомплекс – устойчивое терминологическое сочетание, характерное для выбранной предметной области. Наиболее распространенными структурами здесь являются сочетания существительного с прилагательным, существительного с существительным в косвенном падеже, существительного с другим существительным в качестве приложения. Имеются также многословные термины, иногда состоящие из трех и более слов [3].

074-SidorovaE:_Layout 1 18.05.2008 16:15 Page 478

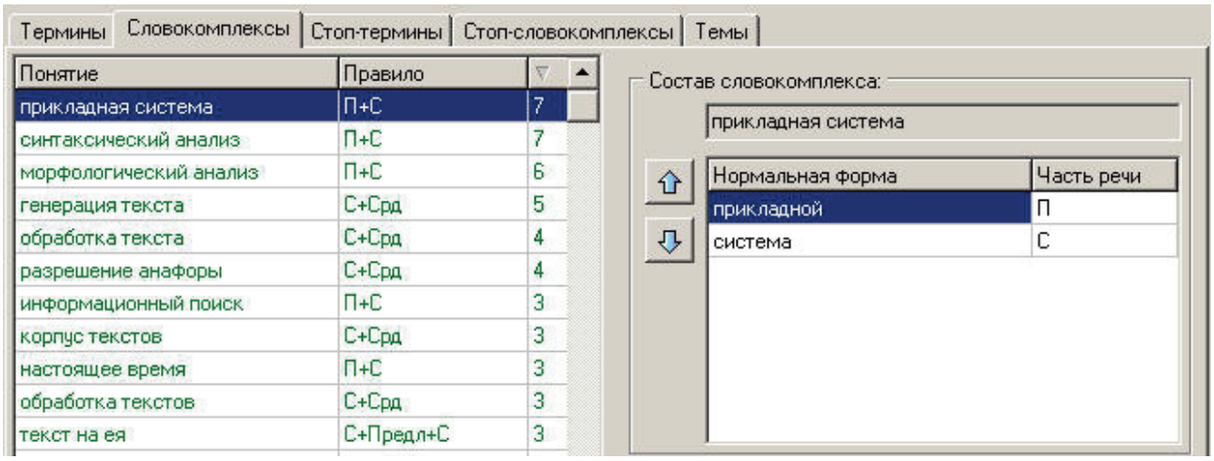


Рис. 2. Фрагмент словаря словокомплексов.

Словарная статья словокомплекса (Рис.2.) содержит следующие терминообразующие признаки:

- нормальная форма,
- список составляющих терминов,
- правило, согласно которому образуется данное сочетание.

Структура словарной статьи словокомплекса фиксирована и не может быть расширена пользователем.

Лексическая конструкция. Под лексической конструкцией понимается несловарная единица, имеющая регулярную структуру, например, номер телефона, дата, инициалы и т.п. Также с помощью лексических конструкций могут описываться специфические термины предметной области, отсутствующие в универсальном словаре русского языка. Для создания словаря лексических конструкций используется технология Alex [4]. Лексическая конструкция в словаре Alex – это множество фрагментов текста произвольной сложности (в общем случае, разрывных), представляющее собой список альтернатив, связываемых с определенной строковой конструкцией.

Отдельной группой стоят формальные сегменты, служащие для сопоставления тексту его жанровой модели, выделения текстовых структур, которые впоследствии могут существенно ограничивать смысл, содержащейся в текстовом фрагменте информации.

2.3. Словари формальных сегментов

Формальный сегмент, по сути, это та же лексическая конструкция, но которая определяет не весь сегмент, а только маркирующие его элементы. Элементы либо задают начало и конец сегмента явным образом, либо неявно через границы другого формального сегмента (в этом случае маркирующий элемент должен быть вложен в сегмент). В качестве маркирующего элемента может выступать либо произвольная строка, либо любой термин, определенный в других словарях, используемых совместно со словарем сегментов.

К терминообразующим признакам сегмента относятся ограничения, на основании которых осуществляется поиск и формирование сегмента:

- single – сегмент не должен пересекаться с сегментами того же типа; частный случай этого ограничения – отсутствие вложенности,
- min – выбираются минимальные из возможных сегменты,
- max – выбираются максимальные из возможных сегменты,
- count – задается максимальное количество сегментов, извлекаемых из текста.

Возможны и другие типы ограничений.

Семантические признаки сегментов находят свое отражение в жанровой модели документа, которая определяется набором сегментов, порядком их следования и вложенностью. Модель документа впоследствии может быть использована для автоматического определения жанра документа.

Примеры жанровых моделей документов можно найти в [4, 5].

074-SidorovaE:_Layout 1 18.05.2008 16:15 Page 479

3. Архитектура словарной подсистемы

Для того, чтобы применять несколько словарей и словарных компонент для решения некоторой задачи, необходим менеджер, осуществляющий управление потоками данных и скрывающий от внешнего пользователя (программы) особенности используемых словарей. Для разных задач менеджер загружает различные сценарии, которые определяют порядок применения словарей, набор входных параметров и схему согласования результатов словарного анализа.

Рис. 3. Архитектура словарной подсистемы

Созданная словарная технология включает словарные компоненты и обработчики, которые обеспечивают с одной стороны, автоматизацию создания и наполнения словарей, с другой стороны, словарный анализ и последующую работы со словарной информацией найденных в тексте терминов (означивание классов термина и его индивидуальных признаков).

Для целей создания словарей были разработаны следующие модули.

- Модуль просмотра конкорданса – позволяет в выбранном корпусе текстов просматривать места встречаемости термина словаря и его контекст.
- Модуль тематизации – обеспечивает анализ текста в различных режимах: наполнение словаря (обучение), ведение статистики встречаемости терминов, классификация на основе статистики. Последовательный анализ текста в разных режимах позволяет поддерживать механизм расширения иерархии классов и «дообучения» словаря.
- Модуль выявления стоп-терминов – позволяет отделить шумовую или общеупотребительную лексику от предметно-зависимой.
- Модуль сборки словокомплексов – извлекает из текста словосочетания по фиксированному набору правил. Основной задачей модуля является выявление наиболее важных терминообразующих синтаксических групп, большинство из которых представляют собой именные группы либо строятся на их основе.

074-SidorovaE:_Layout 1 18.05.2008 16:15 Page 480

- Модуль морфологического анализа Lemmatizer, созданный компанией Диалинг (www.aot.ru), который, в частности, используется для предсказания морфологических признаков незнакомых слов.

Для обеспечения словарной обработки текста дополнительно могут использоваться следующие обработчики:

- Подсистема Alex осуществляет идентификацию и сборку лексических конструкций (текстовых фрагментов), на основе системы правил-шаблонов.
- Модуль исправления ошибок может осуществлять обработку незнакомых слов в режиме словарного анализа. Если слово не прошло словарный контроль, то запускается неточный словарный контроль, реализованный на основе метода хеширования по сигнатуре, предложенного в работе [6]. Модуль использует расширенные словари, содержащие универсальную лексику. Исправление ошибок особенно актуально при обработке неформальных документов: деловой переписке, коротких сообщений, транслитерированного текста.
- Модуль сегментации выполняет первичную и жанровую сегментацию. В процессе первичной сегментации осуществляется разбиение линейного представления текста на строковые объекты, оформленные как сегменты и упорядоченные в соответствии с порядком их встречаемости в тексте. Жанровая сегментация осуществляется после лексического анализа на основе полученных словарных объектов, маркирующих тот или иной жанровый сегмент.

4. Организация доступа к словарной информации

Стандартным способом организации доступа к словарной информации, в частности, к результату словарного анализа текста, является разработка соответствующего программного интерфейса (АПИ). Сам словарный компонент реализуется в виде подключаемой библиотеки, обеспечивающей все операции по работе со словарем, а словари хранятся во внешнем источнике и подгружаются библиотекой во время обработки текста. Отдельно реализуется пользовательский интерфейс (рабочее место) для создания словаря.

Указанный способ, несмотря на его эффективность, может оказаться неудобными по различным причинам, например, использование плохо совместимых сред или языков программирования.

Другим способом доступа является использование универсальных промежуточных форматов представления данных. Одним из самых популярных на сегодняшний день форматов, используемых в разных областях информатики, являются xml-схемы и xml-подобные форматы (rdf, rss, xhtml) [7].

В предлагаемой технологии применяется xml-представление данных, которое позволяет осуществлять следующее.

- Согласование данных. Можно указать семантическую эквивалентность словарных классов или атрибутов из разных словарей (которые могут формироваться с помощью разных программных компонент).
- Расширение содержательного наполнения словарей. Можно объединить элементы разных словарей в одну синонимичную группу (в дальнейшем планируется устанавливать произвольные связи между терминами).
- Более точно специфицировать используемые в словаре признаки для внешней программы или внутреннего обработчика, если необходимая возможность не поддерживается словарным компонентом. Например, в одном из компонентов система семантических классов и рубрик для классификации задаются идентичным образом. Для того, чтобы отличить рубрику от класса во внешнем файле, просто перечисляются все индексы классов.

Таким образом, используемое частичное xml-представление необходимо только для согласования словарей и использования результатов словарного анализа, что позволяет достаточно эффективно использовать данный механизм.

Заключение

На текущий момент одной из самых перспективных задач в данной области является создание комплексного инструмента, позволяющего комплектовать в единый программный продукт различные алгоритмы, реализующие требуемые разработчикам информационных систем этапы анализа текста. Важнейшим компонентом такого комплекса является словарная подсистема, предоставляющая средства для создания, наполнения и использования предметно-ориентированных словарей.

Словари, создаваемы с помощью предложенной словарной технологии, могут поддерживать основные этапы анализа текста: морфологический, синтаксический и семантический, а также классификацию на основе статистики.

074-SidorovaE:_Layout 1 18.05.2008 16:15 Page 481

Планируется проводить дальнейшие исследования проблем и потребностей, возникающих при многоцелевом использовании предметно-ориентированных словарей в практических информационных системах — системах документооборота и специализированных интернет-сервисах.

Список литературы

1. Ершов А.П. К методологии построения диалоговых систем: феномен деловой прозы // Избранные труды. Новосибирск: ВО “Наука”, 1994. С.314-330.

2. Боровикова О.И., Загорулько Ю.А., Сидорова Е.А. Подход к автоматизации сбора онтологической информации для интернет-портала знаний // Труды международного семинара Диалог’2005 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2005. С. 65-70.

3. Большаков И.А. Какие словосочетания следует хранить в словарях? // Труды международного семинара Диалог’2002 по компьютерной лингвистике и ее приложениям. Протвино: 2002. Т.2. С.61–69.

4. Жигалов В.А., Жигалов Д.В., Жуков А.А., Кононенко И.С., Соколова Е.Г., Толдова С.Ю.. Система Alex как средство для многоцелевой автоматизированной обработки текстов // Труды международного семинара Диалог’2002 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2002. Т.2, С.192-208.

5. Кононенко И.С., Сидорова Е.А. Обработка делового письма в системе документооборота // Труды международного семинара Диалог’2002 по компьютерной лингвистике и ее приложениям. М.: Наука, 2002. Т.2. С.299–310.

6. Бойцов Л.М. Использование хеширования по сигнатуре для поиска по сходству // Прикладная математика и информатика. М.: ВМиК МГУ, 2001. № 8. С.135-154.

7. XML // Википедия. http://ru.wikipedia.org/wiki/XML

075-Sokolova_E-Kononenko_I:_Layout 1 18.05.2008 16:17 Page 482

ПРОБЛЕМЫ ОПИСАНИЯ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ В ВИДЕ ОНТОЛОГИИ ДЛЯ ПОРТАЛА ЗНАНИЙ[1]

PROBLEMS OF DESCRIBING COMPUTATIONAL LINGUISTICS IN ONTOLOGY OF A KNOWLEDGE PORTAL

Соколова Е.Г. *(minegot@rambler.ru), Российский государственный гуманитарный университет*
Кононенко И.С. *(irina_k@cn.ru),* ***Загорулько Ю.А.*** *(zagor@iis.nsk.su)*
Институт систем информатики СО РАН, Новосибирск

В этой статье мы обсуждаем проблемы, которые возникли в связи с построением предметной онтологии для научного направления, связанного с компьютерным моделированием языка, преобразований текстов и звучащей речи.

Введение

Существует два подхода к моделированию области исследований – на основе автоматического извлечения терминологии и путем ручного построения онтологий. Разработка портала знаний по компьютерной лингвистике базируется на втором подходе, но предполагает и создание терминологического словаря-тезауруса изучаемой области, поскольку тезаурус (терминологический ресурс, реализованный в виде словаря понятий и терминов со связями между ними [6]) служит инструментом поиска информации в Интернете с целью пополнения информационного содержания портала. При разработке портала по компьютерной лингвистике мы исходили из общей концепции портала знаний, примененной ранее к таким областям знаний, как археология и этнография ([2, 3]).

В основе создания портала знаний по компьютерной лингвистике лежат идеи и принципы, изложенные в работах [7, 8], где в качестве информационной модели портала используется онтология. Под онтологией понимается формальная спецификация концептуальной модели предметной области ([1, 13]).

С содержательной точки зрения, онтология портала представляет систему понятий и отношений, необходимых для описания как научной деятельности и научного знания в целом, так и конкретной области знаний. В связи с этим онтология портала разделяется на метаонтологию и предметную онтологию. В этой статье мы обсуждаем проблемы, которые возникли в связи с построением предметной онтологии для научного направления, связанного с компьютерным моделированием языка, преобразований текстов и звучащей речи. Проблем этих оказалось больше, чем при построении онтологии такой классической гуманитарной науки, как археология. Формулирование проблем скорее способствует выявлению особенностей новой науки, чем помогает созданию онтологии.

Ниже в каждом разделе рассматривается одна конкретная проблема, начиная с названия научного направления и кончая свойствами, связанными с формальными требованиями представления на портале.

Название научного направления

Проблемы начинаются с названия научного направления и установления его границ. Наиболее распространенным названием является термин «компьютерная лингвистика» (КЛ) – калька с английского «Computational Linguistics». КЛ обозначает компьютерные исследования в области лингвистики и ее компьютерных приложений[2]. Термин КЛ вытесняет термин «Прикладная Лингвистика» (ПЛ), содержанием которого является использование моделей языка в прикладных целях. ПЛ у́же, чем КЛ, но шире, чем собственно приложения. Для названия учебной дисциплины в системе высшего образования также используется термин Автоматическая Обработка Естественного Языка (АОЕЯ) (калька с английского «Natural Language Processing» (NLP)). Этот термин не получил у нас широкого распространения. NLP обычно используется для обозначения раздела, посвященного ЕЯ системам, на конференциях по искусственному интеллекту (ИИ). Недостаток этого термина по сравнению с КЛ состоит в том, что он не содержит ключевого для данной области слова «компью-

[1] Работа выполняется при финансовой поддержке РГНФ (проект № 07-04-12149).

[2] Характерно название конференции Диалог, которое в 1995–2001-х годах звучало как «компьютерная лингвистика и ее приложения», а в 2002 г. изменилось на «компьютерная лингвистика и интеллектуальные технологии», включив в сферу КЛ также и приложения, или прикладные системы (applications)

075-Sokolova_E-Kononenko_I:_Layout 1 18.05.2008 16:17 Page 483

тер». Роль компьютера в КЛ отличается от его роли в других науках, где он является только средством усиления и интеллектуализации средств моделирования. В ЛЭС дается определение научной дисциплины Автоматическая Обработка Текста (АОТ) как науки, которая занимается «преобразованием текстов на естественном или искусственном языках с помощью ЭВМ» ([10], стр. 14). Из него следует, что понятие «компьютер» входит в само определение науки. Термин АОТ у́же, чем КЛ, так как не включает исследования по моделированию звучащей речи. В последнее десятилетие появился термин "Linguistic Technology" (LT) и параллельный русский термин «лингвистические технологии».

Из сказанного следует, что наиболее подходящим является термин КЛ, самый широкий, общеупотребительный и содержащий ключевое слово «компьютер». КЛ включает следующие направления: 1) АОТ – первичную для КЛ область преобразования письменных текстов с помощью компьютера, включая приложения; 2) моделирование звучащей речи, включая приложения, 3) мультимодальные приложения, совмещающие обработку текста, звучащей речи и/или изображений; 4) создание языковых ресурсов.

Источники информации

Структуры описания областей знаний обычно создаются в учебниках, которые систематизируют исследования и на этой основе рисуют по возможности полную картину состояния данной области. Они являются выражением опыта авторов учебников, являющихся хорошими специалистами описываемых направлений. Особенность КЛ состоит в том, что она систематизируется не только в виде учебников, но и в виде обзоров, представляющих область КЛ как полную совокупность составляющих ее более узких направлений. Такие обзоры пишутся коллективами авторов – специалистов узких направлений. Для классических наук пишутся только обзоры по узким разделам.

В последние 30 лет исследования по КЛ неоднократно и в разных аспектах описаны в зарубежных учебниках, например, [15, 20, 21, 24], и обзорах [16, 17, 18, 23]. В российских изданиях эта область представлена двумя учебниками – [4] и [12], оба по ПЛ. Российский обзор представлен трехтомником «Искусственный интеллект» [9], содержащим информацию по системам ИИ и КЛ. Западные учебники и обзоры описывают КЛ в разных аспектах – основанный на лингвистике разделяющий теории и приложения подход [23], технологический подход [17], преобладание метода (регулярных выражений) [21], систематизация практических инструментов и методов создания прикладных систем [18].

Изучение учебников и обзоров экспертами предполагает и другие методы сбора информации, необходимые для получения актуальной картины области исследований или ее конкретных разделов. Важнейшим источником информации для эксперта и одновременно источником данных, составляющих информационное пространство портала, является Интернет, где широко представлены информационные ресурсы, отражающие процессы и результаты научных исследований в данной области. К числу наиболее важных результатов в области КЛ относятся лингвистически ориентированные программные средства и лингвистические ресурсы. Информационное наполнение портала такого рода данными невозможно без автоматизации работы эксперта по поиску информационных ресурсов в сети. Эта проблема решается путем создания поискового робота, использующего специальные словари-тезаурусы по КЛ, позволяющие находить, индексировать и предъявлять эксперту странички, относящиеся к данной области, и даже строить гипотезы по их отнесению к тому или иному разделу онтологии КЛ. (см. [5]).

Отличие отечественной парадигмы КЛ от западной

Зарубежные исследования базируются на американских и европейских школах КЛ, для которых характерна разработка формализмов для моделирования лингвистических явлений, а советские и российские исследователи занимались в основном формализацией конкретных языковых явлений. Это противопоставление было отмечено С.А. Шаровым в [14]. Наиболее ярко оно выразилось в создании формализмов для синтаксических анализаторов. Разрабатывались простейшие методы (pattern matching, chunks) и синтаксические формализмы, способные реализовать синтаксический анализ английского языка с постепенным углублением и расширением их возможностей: PSG → трансформационная грамматика (увеличение мощности за счет добавления глубинных структур, в частности, моделирование залоговых трансформаций, преимущественно синтез) → расширенные сети переходов ATN (применимость к анализу) → APSG (возврат к продукциям как преодоление громоздкости ATN) → … ([24]). Процесс завершился изобретением унификационного механизма и созданием унификационных грамматик, которые позволяют моделировать комплексный анализ предложений, используя все формализуемые виды информации, в отличие от предыдущего этапа, основывающегося на сочетаниях синтаксических классов. Из отечественных методов можно отметить два – фильтровый анализ, развивавшийся в работах О.С. Кулагиной

075-Sokolova_E-Kononenko_I:_Layout 1 18.05.2008 16:17 Page 484

и Л.В. Иорданской, и подход на основе продукций, соединяющих различные виды информации (синтагмы) в системе ЭТАП, близкий APSG. Для других реализаций создавались специальные средства, например, язык стандартных операторов (ФРАП).

Российские и западные исследования по грамматике плохо сочетаются в одной онтологии. В частности, возникает вопрос о ценности для российского читателя различных стратегий и алгоритмов реализации PSG, которым было посвящено много западных исследований в 70–80-х годах. С другой стороны, современный этап развития науки, основанный на технологиях, в значительной степени обесценивает большой пласт отечественных исследований, представленных в трехтомнике по ИИ и, частично, в учебниках Гердта и Баранова.

По этому же принципу противопоставлены и лексические исследования. На западе они выразились в создании общедоступных интернет-ресурсов, таких как WordNet, FrameNet и других лексико-семантических баз, например, Corelex – база, описывающая регулярную многозначность существительных английского языка. Появление таких ресурсов вызывало к жизни проекты создания аналогичных ресурсов для других языков и закрепление стандартов. Российские исследования либо сохраняли академический статус, например, Новый объяснительный словарь синонимов русского языка ([11]), либо не приводили к созданию ресурсов, т.е. результаты реализовывались такими средствами, которые исключали их использование другими исследователями и в других системах, например, формализованные описания З.М.Шаляпиной, ЭТАП-2. Приходится сравнивать западные ресурсы с российскими системами и теоретическими описаниями, причем первые отличаются формализованностью и широтой, вторые – глубиной. Появление Национального корпуса русского языка – важный шаг к сближению научных тенденций.

Таким образом, создание онтологии и наполнение портала по КЛ означает необходимость сопоставления, отслеживания соответствий и “наведения мостов“ между подходами, технологиями, методами исследования и терминологией КЛ, традиционно сложившимися в отечественной науке и за рубежом, и представления их в едином “межнациональном“ информационном пространстве. Один из путей решения этой проблемы – связывание результатов и предметов исследования ассоциативными отношениями. Например, лексико-семантический ресурс WordNet и Новый объяснительный словарь синонимов русского языка являются результатами исследований, направленных на разные объекты (разные ЕЯ, английский и русский, и совершенно разные объемы языкового материала: 99808 синсетов версии 1.6 WordNet vs. 354 синонимических ряда Нового словаря), но могут быть связаны по сходству предмета исследования (лексико-семантическая система/подсистема ЕЯ).

Изменчивость направления

В отличие от классических наук, КЛ развивалась по этапам, определенным не внутренней логикой развития, а внешними факторами. КЛ неразрывно связана с компьютером, и ее развитие шло параллельно развитию вычислительной техники, в истории которой выделяется несколько этапов (поколений), связанных с используемыми средствами: 40-е годы 20 века – электронные лампы; 50-е – полупроводниковые приборы; 60-е – интегральные микросхемы; 70–90-е – большие интегральные микросхемы – чипы, которые привели к созданию информационного общества (Интернет, массивы цифровых информационных данных, включая корпуса текстов); и, наконец, происходящий в настоящее время переход к нанотехнологиям.

Каждый из этапов характеризуется определенным качеством развития КЛ и определенным фокусом ее интересов: начальный – МП, 50–60-е годы – осознание непригодности традиционных лингвистических описаний и поиски общей теория языка, пригодной для использования в КЛ, 70–90-е – синтаксические формализмы, семантика, моделирование структуры предложения, 90–2000-е – корпуса, статистика, структура текста, NLG. Проблема состоит в выборе того, что должно быть представлено на портале, потому что наряду со знаниями и информацией, с течением времени сохраняющими свою ценность для исследователей и разработчиков, есть информация, привязанная ко времени и интересная преимущественно в историко-научном плане (для изучающих историю данной дисциплины и преподавателей). Например, Джорджтаунский эксперимент 1954 года имеет значение только как первый в истории опыт по машинному переводу.

Движущей силой КЛ являются актуальные приложения, например, сейчас это – традиционный для КЛ, но сохраняющий свою актуальность, машинный перевод, а также порожденные Интернетом полнотекстовый информационный поиск и извлечение информации.

Особенность теоретической базы

В отличие от классических наук, которые движутся от общего к частному, от общей теории к разработке конкретных ее разделов, развитие теоретической базы КЛ в течение нескольких десятилетий определялось развитием компьютеров и параллельным ему состоянием общественных потребностей. В 70–80-х годах были соз-

075-Sokolova_E-Kononenko_I:_Layout 1 18.05.2008 16:17 Page 485

даны общие модели языка – наиболее известные у нас трансформационная грамматика Н. Хомского и модель «Смысл-Текст» И.А. Мельчука, – а также практически неизвестная у нас системно-функциональная грамматика М. Хэллидея. В настоящее время модель Хомского развилась в лингвистическую теорию «управления – связывания»; модель Мельчука была реализована в виде системы МП ЭТАП-2, систем генерации текстов фирмы CoGenTex и ресурса языковой реализации RealPro. Модель Хэллидея была реализована в виде генератора Penman и затем в виде среды для разработки многоязыковых генераторов текстов KPML. Однако эти системы пока остаются в разряде экспериментальных, не обеспечивая создания эффективных приложений, что является главной целью КЛ. Одна из причин состоит в том, что это – модели языка, а языковые теории и модели не объясняют структуры текста и закономерностей обмена информацией, а именно этот аспект важен для приложений.

Как реакция на углубляющиеся потребности возникли модели структуры текста, в частности, RST и модели, основанные на DRT. Они продвинули понимание структуры текста, связали ее со структурой предложения и коммуникативной целью автора текста. Но они описывают способы подачи информации, а не саму информацию и знания. Пробел в области знаний отчасти компенсируется лингвистическими ресурсами на основе предметных онтологий – WordNet (имеются толкования, иерархия толкований), FrameNet (имеется структура типов процессов и их иерархия). В явном виде знания моделируются в виде общих онтологий, таких как CYC, SUMO, Sowa's ontology.

В 80-х годах возникло новое, эмпирическое, направление в КЛ, основанное на статистических исследованиях на базе корпусов текстов и звуковых корпусов. Его эмпиризм состоит в том, что статистическая модель прямо зависит от материала, на котором она построена и нет явных данных, чтобы объяснить ее действие. (Заметим, что корпуса являются не только основой статистических методов, но и средством верификации созданных вручную ресурсов и теоретических описаний КЛ и традиционной лингвистики.) Статистические методы представляют собой отдельное направление КЛ и описываются в отдельных учебниках, например, [22], но в приложениях в последние 15 лет происходит их слияние с традиционными методами лингвистического анализа «по правилам». Примером этого процесса могут служить вероятностные порождающие грамматики, гибридные методы в системах машинного перевода, использование вероятностных характеристик в алгоритмах морфологического анализа, разрешения неоднозначности, в том числе референциальной, и обработки дискурса.

Процесс интеграции инструментария и методов находит отражение в литературе: [18, 23]. Учитывая это обстоятельство, наиболее общая классификация методов исследования в онтологии КЛ строится не от используемого научного инструментария (логические, статистические, лингвистические), а исходя из предмета/задачи исследования (методы синтаксического анализа, методы разрешения лексической многозначности и т.д.).

В настоящее время, когда в результате развития вычислительной техники внешние факторы, определявшие развитие КЛ в течение нескольких десятилетий исчезли, КЛ впервые получает возможность развиваться по своим внутренним законам и реализовать тягу к научно-обоснованным решениям наряду с эмпирикой.

Мультимодальность

Текст и звучащая речь – основные два вида передачи информации, которые в течение некоторого периода были двумя не зависимыми друг от друга объектами исследования. Речевые технологии развивались как чисто инженерная дисциплина, главным образом направленная на решение практических задач, таких как распознавание голоса диктора. Сейчас в КЛ разрабатываются системы, в которых виды передачи информации на ЕЯ комбинируются: например, в системах общения с компьютерами. Создаются системы, обучающие глухих детей говорить, обучающие читать и понимать прочитанное, системы-гиды и др. Характерно, что в них, кроме текста и звучания, моделируется и зрительная информация. Компьютерная система имеет лицо, мимику, жесты. Последнее вызвало к жизни новое течение КЛ – моделирование невербальной коммуникации, – в котором, кроме жестов, сопровождающих общение людей, исследуются и формализуются языки немых. Таким образом, мультимодальность КЛ проявляется не только в соединении текстовых и речевых технологий, но также текстовых и графических, например, исследования, связанные с поиском на изображении объекта, упомянутого в сопутствующем изображению тексте, см. [19].

Форма «портал знаний», степень детализации и виды размещаемой информации

Форма «портал знаний» накладывает определенные требования на способ представления онтологии КЛ, степень ее детализации и виды размещаемой информации (см. рис.1). Так, формальное деление на классы и экземпляры понятий (объекты), нередко обусловленное дополнительными соображениями удобства навигации и поиска информации на портале, вынудило представить иерархию разделов КЛ с использованием не только отношения таксономии между разделами, представленными как классы, но и отношения партономии между раздела-

075-Sokolova_E-Kononenko_I:_Layout 1 18.05.2008 16:17 Page 486

ми-объектами. Например, показанные на рисунке объекты раздела «Моделирование языка и языковой деятельности» связаны в иерархию отношением партономии: раздел «Речевые технологии» включает «Распознавание речи» и «Синтез речи», а раздел «Автоматическая обработка» текста включает «Понимание текста» и «Генерацию текста».

Существуют трудности, связанные с определением необходимого и достаточного уровня развития/проработки и степени детализации онтологии, которая, не являясь полным справочником по данной предметной области, должна представить и систематизировать ее основные элементы и дать возможность указать информационные источники (ресурсы), которые содержат более подробную информацию.

Возникают также проблемы с представлением «научной деятельности», подробностью отражения конференций и семинаров, персональных данных, а также проектов, не завершившихся общедоступными результатами.

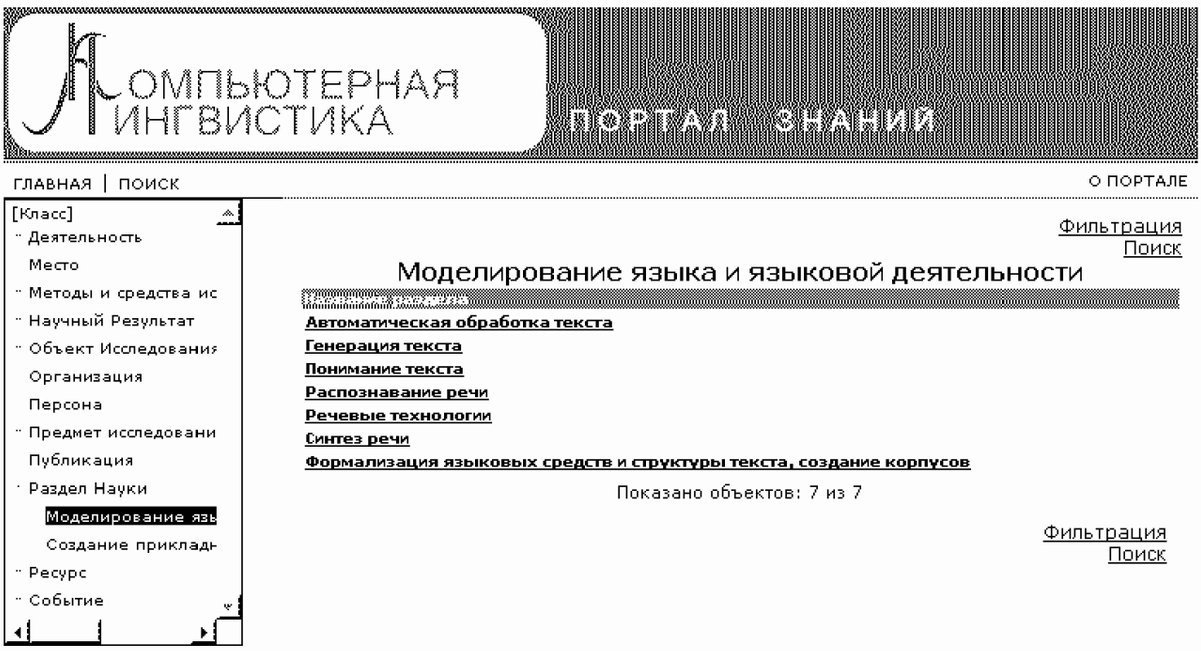


Рис.1. Фрагмент онтологии КЛ на портале знаний

Заключение

Особенности предметной области КЛ, в частности, ее сложность и разветвленность, требуют при разработке онтологии КЛ привлечения квалифицированных экспертов и организации совместной работы специалистов в отдельных направлениях КЛ. По сути, создавая онтологию для портала знаний по КЛ, мы попытались наметить ядро, которое включает наиболее общие классификации и нуждается в развитии и дополнении. В частности, в онтологии в ее текущем виде очень эскизно представлены речевые технологии и очень фрагментарно – обширная область статистических методов, едва намечены методы, связанные с применением нейронных сетей и т.д.

В связи с вышеизложенными проблемами мы нуждаемся в обсуждениях и участии коллег. Поэтому мы сочли возможным представить эскиз портала знаний, создаваемого на основе онтологии КЛ, в Интернете по адресу http://speedy.iis.nsk.su/cl.

Список литературы

1. Александровский Д.А., Кормалев Д.А., Кормалева М.С., Куршев Е.П., Сулейманова Е.А., Трофимов И.В. Развитие средств аналитической обработки текста в системе ИСИДА-Т //Труды 10-й национальной конференции по искусственному интеллекту с международным участием КИИ'2006. М.: Физматлит, 2006. Т. 2. С. 555-563. http://www.raai.org/resurs/papers/kii-2006/doklad/Alexandrovsky.doc

075-Sokolova_E-Kononenko_I:_Layout 1 18.05.2008 16:17 Page 487

2. Андреева О.А., Боровикова О.И., Булгаков С.В., Загорулько Ю.А., Сидорова Е.А., Циркин Б.Г., Холюшкин Ю.П. Археологический портал знаний: содержательный доступ к знаниям и информационным ресурсам по археологии //Труды 10-й национальной конференции по искусственному интеллекту с международным участием КИИ'2006. М.: Физматлит, 2006. Т.3. С.832-840.

3. Археологический портал знаний http://www.sati.archaeology.nsc.ru/classarch2/

4. Баранов А.Н. Введение в прикладную лингвистику. Серия «Новый лингвистический учебник» //М.: Эдиториал УРРС, 2001.

5. Боровикова О.И., Загорулько Ю.А., Сидорова Е.А. Подход к автоматизации сбора онтологической информации для интернет-портала знаний //Труды международной конференции Диалог'2005 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2005. С. 65-70.

6. Журавлев С.В., Добров Б.В. УИС «РОССИЯ». Автоматическое тематическое индексирование полнотекстовых документов //Материалы научно-практической конференции «Проблемы обработки больших массивов неструктурированных текстовых документов», 2001. http://www.fep.ru/text/dataarrays03.html

7. Загорулько Ю.А., Боровикова О.И. Методологические проблемы построения и использования онтологий в портале научных знаний //Труды IX Международной конференции «Проблемы управления и моделирования в сложных системах». Самара: Самарский Научный Центр РАН, 2007. С. 447-454.

8. Загорулько Ю.А., Боровикова О.И., Загорулько Г.Б. Организация содержательного доступа к информационным ресурсам на основе онтологий //Труды 9-й Всероссийской научной конф. "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" RCDL'2007. Переславль-Залесский, 2007. Т. 1. С. 217-224.

9. Искусственный интеллект. Справочник в 3 кн. . Под ред. Попова Э.В //М.: Радио и связь, 1990.

10. Лингвистический энциклопедический словарь. Под ред. Ярцевой В.Н. //М.: Советская энциклопедия, 1990.

11. Новый объяснительный словарь синонимов русского языка. Под общим руководством акад. Апресяна Ю.Д. //Второе издание. Москва, Вена: Языки славянской культуры: Венский славистический альманах, 2004.

12. Прикладное языкознание. Учебник. (ред. Гердт А.С.) //СПб., 1996.

13. Смирнов С.В. Онтологии в прикладных интеллектуальных системах: прагматический подход //Труды 9-й национальной конференции по искусственному интеллекту с международным участием КИИ'2004. М.: Физматлит, 2004. Т.3. С.1059-1067.

14. Шаров С.А. Средства компьютерного представления лингвистической информации //1996. http://www.ksu.ru/eng/science/ittc/vol000/002/

15. Allen J. Natural Language Understanding //Benjamin Cummings, 1995.

16. Butter, C.S. (ed.) Computers and written text //Blackwell, Oxford (UK) and Cambridge (USA), 1992.

17. Cole Ronald (ed.) Survey of the state of the Art in Human Language Technology //1996. (http://cslu.cse.ogi.edu/HLTsurvey/).

18. Dale R., Moisi H., Somers H. (eds.) Handbook of Natural Language Processing //Marcel Dekker, New York, 2000.

19. Deschacht K., Moens M-F. Text analysis for automatic image annotation //The 45th Annual meeting of the Association for Computational Linguistics, Prague, June 2007. http://acl.ldc.upenn.edu/P/P07/P07-1126.pdf

20. Grishman Ralph. Computational linguistics. An introduction //Cambridge, 1986.

21. Jurafsky Daniel, Martin James H. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics //Prentice Hall, 2000.

22. Manning Ch., Schütze H., Foundations of Statistical Natural Language Processing //MIT Press. Cambridge, MA, May 1999. http://nlp.stanford.edu/fsnlp/

23. Mitkov Ruslan (ed.) The Oxford handbook of computational linguistics //N.Y.: Oxford university press, 2003.

24. Salton Gerard. Automatic text processing. The transformation, analysis, and retrieval of information by computer //Addison-Wesley Publishing Company, Inc., 1989.

076-Bogdanova2:_Layout 1 18.05.2008 16:17 Page 488

ЗВУКОВОЙ КОРПУС РУССКОГО ЯЗЫКА ПОВСЕДНЕВНОГО ОБЩЕНИЯ «ОДИН РЕЧЕВОЙ ДЕНЬ»: КОНЦЕПЦИЯ И СОСТОЯНИЕ ФОРМИРОВАНИЯ

SPEECH CORPUS OF THE RUSSIAN EVERYDAY COMMUNICATION «ONE SPEAKER'S DAY»: BASIC CONCEPTION AND CURRENT STATE

Степанова С.Б. *(stsvet_2002@mail.ru),* ***Асиновский А.С.*** *(a.s.asinovsky@gmail.com),* ***Богданова Н.В****. (nvbogdanova_2005@mail.ru),* ***Русакова М.В****. (mvrusakova@gmail.com),* ***Шерстинова Т.Ю.*** *(sherstinova@gmail.com)*
Факультет филологии и искусств Санкт-Петербургского государственного университета

В докладе рассматриваются методические принципы создания звукового корпуса русского языка повседневного общения «Один речевой день», даны правила первичной обработки речевого материала и описание специализированной базы данных, приводятся сведения о текущем состоянии формирования корпуса.

1. Введение

Начиная с 1990х гг. во многих странах мира создаются национальные корпусы спонтанной речи. Впервые подобный звуковой корпус, записанный от демографически сбалансированной выборки информантов, был создан в рамках Британского национального корпуса. Для русской речи представительного корпуса языка повседневного общения ранее не существовало, хотя звуковые базы данных создаются в разных научных коллективах на протяжении уже как минимум 40 лет1.

Несмотря на известный опыт в изучении разговорной речи, многое здесь еще остается неизвестным. Например, сколько различных словоформ (морфем, высказываний) человек порождает и воспринимает в течение часа, дня, месяца; сколько разных языковых единиц он употребляет в разные периоды времени или какова общая продолжительность звучания его речи за те же периоды.

Отсутствие подобной лингвистической информации вполне объяснимо: проведение такого рода исследований исключительно трудоемко, технические возможности для их проведения появились относительно недавно. Несмотря на то что сейчас такие возможности существуют, эти исследования сугубо инновационны, поскольку в них должны быть вовлечены довольно большие профессиональные научные коллективы.

В последние десятилетия собран значительный объем данных, связанных с естественной спонтанной речью на русском языке. К сожалению, разные исследователи накапливают материал с применением различных методик и в соответствии со своими задачами. В результате собранные ресурсы оказываются неоднородными, разрозненными и требуют больших усилий для формирования из них единого представительного корпуса. Следовательно, создание полноценной описательной базы русской устной разговорной речи является необходимым условием развития тех областей лингвистики, которые обращены к человеку и его речевому поведению.

2. Концепция и принципы создания звукового корпуса «Один речевой день»

2.1. Методические предпосылки

Приоритетная задача данного исследования заключается в том, чтобы получить записи русской спонтанной речи в *естественных* условиях. Это значит, что, во-первых, ничто не должно влиять на особенности речевого поведения говорящего в конкретных речевых ситуациях. Например, речевая коммуникация во время завтрака должна осуществляться в обычных для каждого информанта условиях: в том же помещении и с теми же коммуникантами, как обычно, с тем же уровнем шума (при открытом или закрытом окне, шуме холодильника и

[1] Подробнее об этом см. Полевая лингвистическая практика. Учебно-методический комплекс сложной структуры. Часть 1. Теоретические основы и методика сбора лингвистических данных для представления их в речевом корпусе русского языка / Ред. *Асиновский А. С., Богданова Н. В.* СПб., 2007. В качестве примера таких баз можно назвать проект «Отчеты детей об их сновидениях» (руководители А. А. Кибрик и В. И. Подлесская)», а также устную часть Национального корпуса русского языка.

076-Bogdanova2:_Layout 1 18.05.2008 16:17 Page 489

т. п.). Во-вторых, информант реализует свое речевое поведение в стандартных для него ситуациях, не меняя ради записи их репертуар и продолжительность. Например, если во время завтрака он привык читать газету, то не следует отказываться от этой привычки и вступать в несвойственную для этого момента коммуникацию с членами семьи или приглашать незапланированных гостей, чтобы увеличить объем речевой продукции во время записи. В то же время на начальном этапе исследования предполагается выбор дней, в которые проводится запись. Например, предпочтительнее осуществлять запись в обычный день, а не тогда, когда информант совершает необычные для себя действия (едет на экскурсию, отсутствует дома целый день из-за сезонного аврала на работе и т. п.).

2.2. Техническое обеспечение сбора материала

Запись проводится с использованием диктофона, который, предварительно настроив, информант закрепляет на себе стационарно. Такой режим записи неминуемо приводит к тому, что качество собранного материала получается неоднородным. В нашем случае запись осуществлялась на цифровые диктофоны *Olympus WS-320M*, обеспечивающие более 35 часов качественной записи.

Тем не менее, относительно низкий уровень качества такой записи по сравнению со студийными является неустранимым следствием полевой работы в условиях естественного эксперимента с речью и речевым поведением человека.

2.3. Отбор информантов

На данном этапе исследования не ставится задачи описать функционирование русского языка во всем многообразии его проявлений. Исследованию подвергается только одна из форм его бытования - речь наивных носителей языка, для которых русский является родным, жителей города, расположенного в сфере господства литературного языка, не подвергавшегося никаким мощным диалектным или иноязычным влияниям, города с полипрофессиональным населением, без сдвигов в возрастном и гендерном распределении. Практически идеальным образцом такого города является Санкт-Петербург. Именно поэтому отбор информантов проводится среди жителей Санкт-Петербурга.

Особенности данного исследования предполагают, что в роли информантов не должны выступать люди, привыкшие профессионально следить за своей речью, так как соответствующие навыки могут существенным образом повлиять на качество речевой продукции.

2.4. Работа с информантами

Работа с информантами в основном складывается из двух этапов: 1) обеспечение естественности коммуникативного поведения информантов; 2) обеспечение качества записи, необходимого для дальнейшего анализа полученного материала.

Выполнение первой задачи является очень трудным. Пилотный эксперимент, в котором информантами стали сами члены исследовательского коллектива, показал, что, несмотря на высокую мотивированность получения адекватных результатов, человек практически не может «забыть о микрофоне», если знает, что его знакомые (в данном случае коллеги) станут свидетелями его коммуникативных контактов с другими людьми, в особенности с близкими родственниками. Для получения максимально естественных записей необходимо проведение сбора материала в условиях полной анонимности. Для достижения этого была разработана следующая процедура. В исследовании принимает участие сотрудник, психолингвист по специальности, не являющийся членом рабочей группы. Он обращается к потенциальным информантам, например, к работникам какого-то предприятия. Проводя инструктирование, он дает каждому гарантию того, что сам не будет работать с полученными записями. Впоследствии сборщик передает записанный материал исследовательскому коллективу. Информанты не сообщают своих имен и фамилий, но заполняют специально разработанную анкету, где есть вопросы о возрасте, специальности, месте рождения и т. п. В результате исследователи работают с речевой продукцией людей, не только абсолютно незнакомых, но и никогда им не встречавшихся. Оборотной стороной данной процедуры является то, что информанты оказываются вне всякого контроля со стороны лингвиста-профессионала, что заранее предполагает довольно высокий уровень брака, большое количество записей, непригодных для дальнейшей обработки.

Выполнение второй задачи достигается инструктированием информантов по поводу использования диктофона.

076-Bogdanova2:_Layout 1 18.05.2008 16:17 Page 490

2.5. Первичная обработка речевого материала

Обработка материала осуществляется квалифицированными исследователями-лингвистами. Первичная обработка – это предварительное описание материала и его орфографическая расшифровка. В исследованиях разговорной речи расшифровщиками чаще всего являются сами участники разговора. Именно они способны наиболее адекватно расшифровать фрагменты, характеризующиеся плохой разборчивостью, и описать экстралингвистический фон, на котором проходит общение. Однако в нашем исследовании расшифровщики не являются не только участниками, но и свидетелями коммуникативного поведения информантов. Расшифровка речи не участниками общения естественным образом приводит к тому, что значительная часть информации утрачивается. Это должно рассматриваться как своеобразная «плата» за естественность собранного материала.

Для орфографической расшифровки речевого материала экспертам были предложены следующие *правила*.

При неправильной постановке ударения или возможной его вариативности гласный выделяется с помощью большой буквы: *складЫ, творОг*.

Фонетическая транскрипция не включается в орфографическую расшифровку. В некоторых особо очевидных отклонениях от кодифицированного литературного языка в графе «комментарии» делались пометки типа: [чек] (*человек*), [грю] (*говорю*) и т. п.

Разрядкой передается замедление темпа.

Ремаркой [*нрзб*.] обозначается неразборчивость слова или части записи; если фрагмент неразборчивого участка больше слова, указывается время его начала и конца.

Скандирование, затягивание гласных и согласных передается с помощью дефисов: *Ну-у! Не по-ни-ма-ю*!

Некоторые явления «неканонической фонетики» условно передаются следующими орфограммами:

угу - утвердительное междометие, произносимое с закрытым ртом;

гм – произносимое с закрытым ртом звуки;

3) *не-а* – отрицание, вторая часть которого может произноситься с твердым приступом (гортанной смычкой);

4) *не-у* – междометие отрицания, произносимое с закрытым ртом;

5) *м-м, э-э, а-а* – заполнение хезитационных пауз;

6) *М*? – переспрос с закрытым ртом.

При членении речевого текста используются следующие знаки:

1) / - перцептивная межсинтагменная пауза (при этом может не быть чисто физического перерыва в звучании) – там, где ее наличие ощущается как нормативное: *Я вчера ходила в кино / и там Женю встретила //*;

2) // - реальная пауза достаточно большой длительности. Если она находится после отрывка с интонацией завершения, считаем её заменяющей знак точки и следующую реплику пишем с большой буквы: *Я вчера ходила в кино // Там Женю встретила //*;

3) (…) - хезитационная пауза: реальная пауза там, где её наличие ощущается как ненормативное: *Я вчера ходила …в кино / там Женю встретила //*;

4) ! ? – знаки для передачи восклицательных и вопросительных реплик. Они заменяют знак //.

Для систематизации корпуса и представления результатов обработки данных была разработана специализированная база данных.

3. Специализированная база данных SpeechDay

Звуковой корпус «Один речевой день» состоит из двух модулей: массива звуковых файлов и базы данных SpeechDay. Последняя представляет собой реляционную базу данных, разработанную в формате MS Access 2003. На настоящий момент (версия 1.2) она состоит из 7 таблиц, которые можно условно разделить на 2 группы: фактические данные и результаты научно-исследовательской работы и их интерпретация. Некоторые таблицы содержат «смешанные» данные.

Группа 1

Таблица 1 – *Informants*: фактические данные обо всех базовых информантах, полученные из анкет, заполненных самими информантами. В связи с тем что ответы на многие вопросы не являлись для обязательными, поля этой таблицы заполнены не полностью.

076-Bogdanova2:_Layout 1 18.05.2008 16:17 Page 491

Поле	Описание
N	порядковый номер информанта в базе данных;
FIO	фамилия-имя-отчество информанта или его псевдоним (внутренняя информация разработчиков корпуса);
BInf	код информанта (И1, И2, etc.);
Gender	пол информанта (Ж/М);
Age	возраст информанта (число полных лет) - точно или приблизительно;
PBirth	место рождения информанта (город, регион и т. п.);
SClass	социальное происхождение (напр., профессия родителей и т. п.);
Educ	образование информанта (высшее, среднее и т. п.);
Qual	квалификация информанта (по диплому, свидетельству и т. п.);
Prof	фактическая профессия или характер деятельности информанта на момент записи;
Nat	национальность информанта и его родителей (по отдельности);
InfComments	комментарий относительно типа личности и речевых особенностей информанта;
NFiles	количество звуковых файлов, полученных от информанта;
QFiles	качество записанных звуковых файлов;
TTime	общее время записи;
RTime	полезное время записи (относительно разборчивая речь);
RecComments	комментарий относительно звукозаписей, полученных от данного информанта;
RName	фамилия-имя-отчество исследователя, выполнявшего расшифровку.

Таблица 2 – *Communicants*: фактические данные обо всех коммуникантах, также полученные из анкет, заполненных информантами.

Поле	Описание
BInf	код информанта (И1, И2, etc.);
OInf	код коммуниканта (А1, F2, etc.);
FIO	фамилия-имя-отчество коммуниканта или его псевдоним (внутренняя информация разработчиков корпуса);
Relation	отношение коммуниканта к информанту (напр., мать, друг, продавец, etc.);
Gender	пол коммуниканта (Ж/М);
Age	возраст коммуниканта (число полных лет) - точно или приблизительно;
PBirth	место рождения коммуниканта (город, регион и т. п.);
SClass	социальное происхождение (напр., профессия родителей и т. п.);
Educ	образование коммуниканта (высшее, среднее и т. п.);
Qual	квалификация коммуниканта (по диплому, свидетельству и т. п.);
Prof	фактическая профессия или характер деятельности коммуниканта на момент записи;
Nat	национальность информанта и его родителей (по отдельности);
Comments	комментарий относительно типа личности и речевых особенностей коммуниканта;
SpeechSample	имя звукового файла или имя исходного звукового файла и точный адрес метки начала речи;
RQuality	качество записи по 5балльной системе (5 - максимум).

Таблица 3 – *SoundFiles*: описание исходных звуковых файлов, полученных от информантов - информация о длительности каждого звукового файла, о длительности частей, которые поддаются расшифровке (полной или частичной), о длительности частей, которые не содержат речи или не поддаются расшифровке.

Поле	Описание
BInf	код информанта (И1, И2 и т. п.);
SFile	имя звукового файла (напр., w1.wav);
TTime	общее время звучания;
RTime	полезное время звучания;
Comments	комментарий к файлу;
OverView	наличие (галочка) / отсутствие «скоростной» расшифровки;
Decoding	наличие (галочка) / отсутствие детальной расшифровки;
TDecoding	временна́я продолжительность детальной расшифровки.

076-Bogdanova2:_Layout 1 18.05.2008 16:17 Page 492

Таблица 4 – *Epizods*: описание основных эпизодов речевого дня, полученных в результате прослушивания звукозаписей экспертами, информация об участниках разговора, о времени начала и конца каждого сюжета, о его теме, о месте и времени разговора. На первом этапе заполнения базы разбивка на эпизоды выполнялась относительно произвольно, по усмотрению эксперта. В настоящий момент ведется работа по стандартизации (нормализации) этих данных.

Поле	Описание
BInf	код информанта (И1, И2, etc.);
SFile	имя звукового файла (напр., w1.wav);
NScene	порядковый номер фрагмента/эпизода;
SceneName	название эпизода (напр., «говорят о погоде»);
STime	начало звучания;
FTime	полезное время звучания;
Decoding	наличие (галочка) / отсутствие детальной расшифровки;
Speakers	коммуниканты, участвующие в разговоре;
Time	точное или примерное время разговора (10 часов утра / утро и т. п.);
Place	место, где происходит разговор (дом, офис, etc.);
Overview	описание ситуации (что происходит / *нрзб* / шум);
Comments	комментарий к данному фрагменту звукозаписи.

Таблица 5 – *Decoding*: содержит подробную орфографическую расшифровку отдельных эпизодов, выполненную экспертами. Планируется заполнение этой таблицы для всего корпуса.

Поле	Описание
BInf	код информанта (И1, И2, etc.);
SFile	имя звукового файла (напр., w1.wav);
NScene	порядковый номер фрагмента/эпизода;
STime	начало фрагмента;
Speaker	говорящий (код информанта или коммуниканта);
Speech	расшифровка речи или описание (не)вербальной ситуации (напр., шум / *нрзб* и т. п.);
Comments	комментарий;
OComments	другие комментарии (лексический, грамматический и пр.).

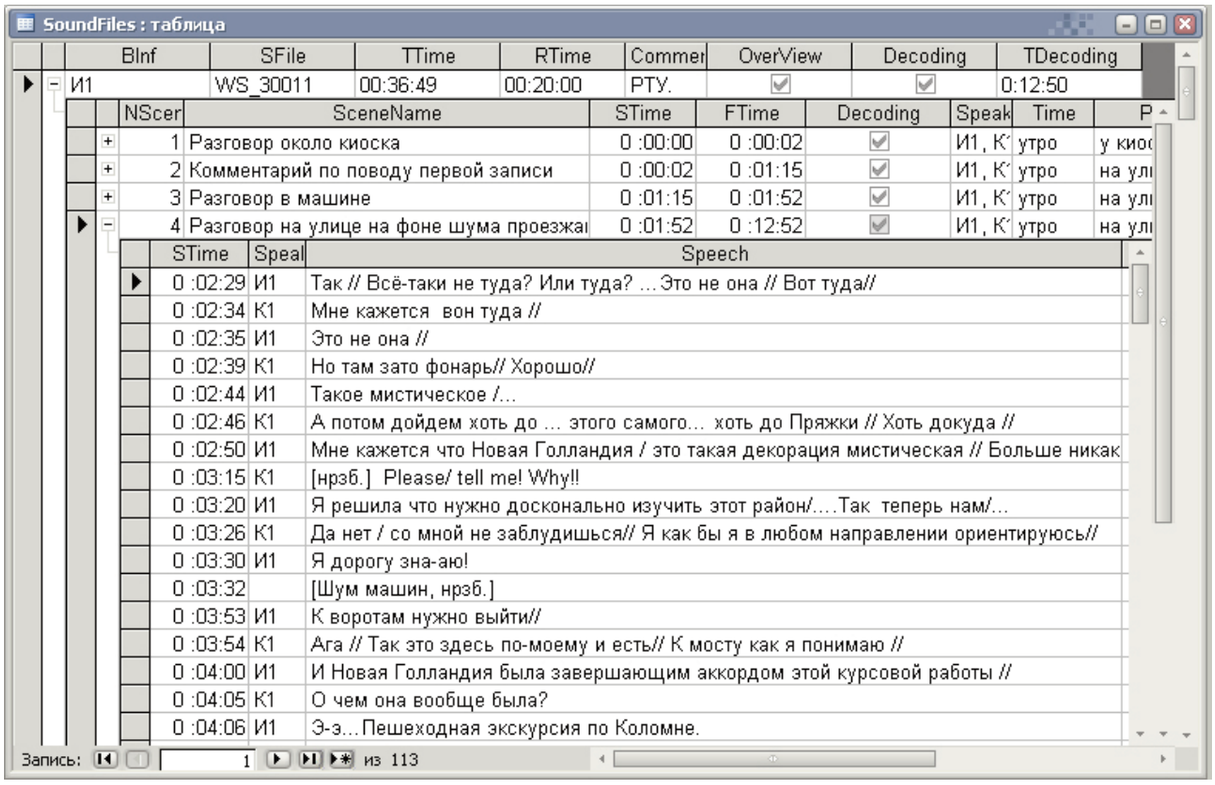
Группа 2 представлена пока лишь двумя таблицами.

Таблица 6 – *InformantsSocial*: имеет ту же структуру, что и *Informants*, однако данные, представленные в ней, – это результат субъективной оценки информанта исследователем, который работал с соответствующим материалом. Следует заметить, что до заполнения этой таблицы исследователи не были знакомы с фактическими данными таблицы 1.

Таблица 7 – *InformantsPsycho*: содержит данные, которые можно охарактеризовать как «психологический портрет» информанта. Таблица заполнялась исследователями, работавшими с записями.

Поле	Описание
BInf	код информанта (И1, И2, etc.);
Nev	невротичность;
SAgr	спонтанная агрессивность;
Depr	депрессивность;
Razd	раздражительность;
Obsch	общительность;
Uravn	уравновешенность;
RAgr	реактивная агрессивность;
Zast	застенчивость;
Otkr	открытость;
Extr	экстраверсия / интроверсия;
Emot	эмоциональная лабильность;
Mask	маскулинность / феменизм;
Esse	словесный портрет информанта, написанный исследователем.

076-Bogdanova2:_Layout 1 18.05.2008 16:17 Page 493



SoundFiles : таблица

BInf	SFile	TTime	RTime	Commer	OverView	Decoding	TDecoding
И1	WS_30011	00:36:49	00:20:00	РТУ.	☑	☑	0:12:50

NScer	SceneName	STime	FTime	Decoding	Speak	Time	P
1	Разговор около киоска	0 :00:00	0 :00:02	☑	И1, К'	утро	у кио
2	Комментарий по поводу первой записи	0 :00:02	0 :01:15	☑	И1, К'	утро	на ул
3	Разговор в машине	0 :01:15	0 :01:52	☑	И1, К'	утро	на ул
4	Разговор на улице на фоне шума проезжа	0 :01:52	0 :12:52	☑	И1, К'	утро	на ул

STime	Speal	Speech
0 :02:29	И1	Так // Всё-таки не туда? Или туда? ...Это не она // Вот туда//
0 :02:34	К1	Мне кажется вон туда //
0 :02:35	И1	Это не она //
0 :02:39	К1	Но там зато фонарь// Хорошо//
0 :02:44	И1	Такое мистическое /...
0 :02:46	К1	А потом дойдем хоть до ... этого самого... хоть до Пряжки // Хоть докуда //
0 :02:50	И1	Мне кажется что Новая Голландия / это такая декорация мистическая // Больше никак
0 :03:15	К1	[нрзб.] Please/ tell me! Why!!
0 :03:20	И1	Я решила что нужно досконально изучить этот район/....Так теперь нам/...
0 :03:26	К1	Да нет / со мной не заблудишься// Я как бы я в любом направлении ориентируюсь//
0 :03:30	И1	Я дорогу зна-аю!
0 :03:32		[Шум машин, нрзб.]
0 :03:53	И1	К воротам нужно выйти//
0 :03:54	К1	Ага // Так это здесь по-моему и есть// К мосту как я понимаю //
0 :04:00	И1	И Новая Голландия была завершающим аккордом этой курсовой работы //
0 :04:05	К1	О чем она вообще была?
0 :04:06	И1	Э-э...Пешеходная экскурсия по Коломне.

Запись: 1 из 113

Рис. 1. Фрагмент заполненной базы данных (таблицы SoundFiles > Epizods > Decodings)

4. Формирование звукового корпуса «Один речевой день»

Осенью 2007 г. при поддержке гранта РГНФ (проект № 07-04-94515е/Я *Звуковой корпус русского языка повседневного общения «Один речевой день»*) была осуществлена первая серия звукозаписей. Для этого была отобрана группа информантов из 30 человек, представляющих разные социальные и возрастные слои населения Санкт-Петербурга. Информанты после подробного инструктажа осуществили звукозаписи своих речевых контактов в течение одного дня, а также заполнили социологические анкеты и прошли психологическое тестирование.

Полученные записи представлены в 266 файлах. Прослушано и частично расшифровано 202 файла (часть файлов была забракована из-за отсутствия в них речи). Общая длительность записанного материала – 195 часов, из них 134 часа содержат речь, вполне пригодную для дальнейшего анализа.

Кроме речи информантов, записана также речь 520 их коммуникантов. Среди них – люди самого разного возраста (от 3 до 68 лет), разных специальностей (продавцы, кондукторы, менеджеры, преподаватели, врачи, библиотекари, компьютерщики и др.), а также студенты и курсанты, состоящие с информантами в родственных, дружеских, производственных или иных отношениях. В материалах представлены разнообразные жанры и стили речи: деловой разговор с коллегами, разговор по телефону, чтение лекции, проведение практических занятий по иностранному языку, общение с друзьями и родными во время прогулок или вечеринок, за ужином, за завтраком и т. п. Темы разговоров также разнообразны: от обсуждения со стоматологом проблем с зубами до разговоров о религии, о жизни и смерти. Записи были сделаны в домашних условиях, в транспорте, на улице, в университете, в военном училище, в кафе, в магазине, в парке аттракционов.

Из записей выделено и проаннотировано 2202 эпизода, из них подробно расшифровано 134.

Подробно расшифровано более 4500 реплик, 39300 словоупотреблений, 220300 знаков (без символов «/» и «//»), что составляет примерно 5,5 печ. листов.

Опыт использования базы данных SpeechDay показал, что на следующем этапе целесообразно оптимизировать структуру базы данных, осуществив нормализацию целого ряда параметров. Наиболее сложными для нормализации представляются названия эпизодов. Планируется разработка более удобного пользовательского интерфейса и многоуровневой поисковой системы, а также обеспечение доступа к соответствующему звуковому файлу непосредственно из среды базы данных.

076-Bogdanova2:_Layout 1 18.05.2008 16:17 Page 494

Последующие этапы обработки материала будут представлять собой его описание на разных уровнях и с разной степенью подробности в соответствии с различными задачами интегрального описания речи. Планируется расширение речевого материала – получение звукозаписей от новых информантов, оптимизация специализированной базы данных и ее преобразование в мультимедийную информационную систему. Полная реализация проекта будет иметь важное значение как для решения фундаментальных научных задач (изучения русской спонтанной речи на разных лингвистических уровнях, исследования реальных коммуникативных ситуаций и сценариев, построения интегральной модели языка повседневного общения), так и для решения актуальных прикладных задач в области речевых технологий (в первую очередь для синтеза и распознавания русской речи).

077-Gerassimenko_1:_Layout 1 18.05.2008 16:18 Page 495

К ВЗАИМОДЕЙСТВИЮ КОМПЬЮТЕРА И ЧЕЛОВЕКА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ*

TOWARDS HUMAN-COMPUTER INTERACTION IN NATURAL LANGUAGE

Странdсон К. *(krista.strandson@ut.ee),* ***Герасименко О.*** *(olga.gerassimenko@ut.ee),*
Кастерпалу Р. *(riina.kasterpalu@ut.ee),* ***Койт М.*** *(mare.koit@ut.ee),* ***Ряэбис А.*** *(andriela.raabis@ut.ee)*
Тартуский университет, Эстония

В статье рассмотрена структура естественных справочных диалогов для разработки эстонской диалоговой системы, общающейся с пользователем на естественном языке. На материале Эстонского диалогового корпуса проанализированы языковые характеристики информационных запросов клиента и ответов информатора. Результаты исследования будут использованы в двух диалоговых системах, находящихся в стадии разработки.

1. Введение

Диалоговые корпуса, состоящие из записей естественных разговоров, представляют собой хорошую основу для изучения культурно обусловленных стратегий и методов, используемых собеседниками для достижения целей в разговоре. Наиболее естественным и распространенным способом общения является (устный) диалог. Компьютерная система, претендующая на роль участника диалога, т.е. интеллектуального агента, должна понимать и генерировать единицы естественного языка, а также управлять диалогом наравне с собеседником. Хотя и существует предположение, что люди общаются с компьютерной системой иначе, чем с другим человеком [1], мы убеждены, что основополагающие принципы ведения диалога на естественном языке нужно искать в естественном диалоговом поведении, и лишь затем анализировать специфику общения человека и компьютера.

Существует ряд диалоговых систем на естественном языке, выполняющих информационные запросы пользователя или определенные практические задания [2], [3]. В моделировании диалогового поведения важную роль играют устные диалоговые корпуса (например, в проектах Switchboard, Verbmobil, BNC). Телефонные службы, основанные на устных диалоговых системах, широко внедряются в областях информационного поиска и практических операций, включая системы транспортных расписаний, справочные и технические службы, системы управления интеллектуальным домом и системы навигации.

Наша цель – разработать диалоговую систему (ДС) для устной коммуникации на эстонском языке. Мы изучаем естественные справочные диалоги, чтобы выяснить, как клиент формулирует запрос, какие языковые средства используют участники диалога, какова структура устного институционального диалога и что должна учитывать система, следующая правилам естественной коммуникации. Исследование естественных диалогов важно и для автоматического разпознавания речи: описание частотных шаблонов облегчает расшифровку беглой речи. Схожим образом, описание речевого поведения информатора помогает смоделировать естественное, лаконичное и практичное речевое поведение ДС.

На материале русских справочных диалогов структуру и динамику изменения текущего сознания коммуникантов исследовал А.Е. Кибрик [11]; наше исследование сосредотачивается на эксплицитных деталях речевого взаимодействия, однако мы используем близкий подход.

В разделе 2 мы представляем обзор эмпирического материала – Эстонского диалогового корпуса и типологии диалоговых актов, используемой для разметки корпуса. Разделы 3 и 4 посвящены анализу корпуса и соответственно рассмотрению запросов клиента и ответов информатора. В разделе 5 рассмотрены результаты корпусного анализа, которые будут использованы в двух экспериментальных ДС. Раздел 6 содержит выводы.

2. Эмпирический материал

Наше исследование основано на Эстонском диалоговом корпусе (EDiC, http://math.ut.ee/~koit/Dialoog/EDiC.html). Корпус состоит из записей и расшифрованных текстов более 1100 аутентичных устных диалогов, в т.ч. около 1000 телефонных диалогов. В корпусе размечены диалоговые акты (функции диалогических реплик). Используемая нами типология диалоговых актов основана на

* Работу поддерживает Эстонский научный фонд (грант № 7503) и Министерство просвещения Эстонии.

077-Gerassimenko_1:_Layout 1 18.05.2008 16:18 Page 496

конверсационном анализе (КА) [4], который исследует техники, используемые людьми в социальном взаимодействии. Согласно КА, некоторые диалоговые акты образуют смежные пары, в которых появление первого элемента обуславливает появление второго элемента (так, информационный запрос требует ответа). Если второй элемент не следует непосредственно за первым (например, вместо ответа собеседник задает уточняющий вопрос), его ожидание сохраняется на протяжении вставной секвенции. Компьютер как участник диалога должен быть способен отличить первый элемент смежной пары (ожидающий ответа) от второго элемента, а также от реплик, не ожидающих специфических ответных действий (например, реплик обратной связи: *так/ясно*).

Для описания поведения клиента и информатора в информационных диалогах из корпуса были отобраны 60 случайных звонков в справочную службу. Информационные запросы клиента имеют форму вопроса (специального, напр. *Когда идет последний поезд на Тарту?* или общего, напр. *Вы не подскажете мне последний поезд на Тарту?*) или директива (напр., *Подскажите мне последний поезд на Тарту*). Случаи использования общих и специальных вопросов мы рассматривали в ряде предыдущих статей [5], [6]. В этой статье мы рассмотрим смежные пары директивов и реакций на них («директив – выполнение директива»). Директивный запрос выражает желание или потребность говорящего получить некую информацию. Типичными реакциями на запрос являются предоставление информации или сообщение об отсутствии информации (прим. 1, C – клиент, I – информатор; транскрипция КА).

(1)
C: öelge=palun: `pensioniameti `telefoni (.) .h `number (.) `Tartus.
скажите, пожалуйста, номер телефона пенсионного департамента в Тарту ЗАПРОС
I: ee `number on `seitse=neli=`neli?
ээ, номер семь четыре четыре.. ИНФОРМАЦИЯ

Рассмотрим, какие диалоговые акты используются для оформления запросов и ответов (табл.1).

ЗАПРОСЫ КЛИЕНТА	# ЗАПРОСЫ	ОТВЕТЫ ИНФОРМАТОРА	# ОТВЕТЫ
Запрос	52	Предоставление информации Отсрочка Отсутствие информации	129 31 4
Запрос + Смена темы	2	Предоставление инф.	2
Предоставление инф. + Запрос	1	Предоставление инф.	1
Согласие + Запрос	2	Предоставление инф.	2
Утвердительный ответ + Запрос	1	Предоставление инф.	1
Всего	58		170

Таблица 1. Запросы и ответы в справочных диалогах

Ответных реплик в три раза больше, чем реплик с запросами, так как первой реакцией на запрос обычно является сообщение об отсрочке (информатору нужно время, чтобы свериться с базой данных). Кроме того, информатор часто разбивает ответ на блоки и произносит в несколько реплик (репликой в традиции КА считается поток речи между сменами говорящих (прим. 2), могущий содержать несколько диалоговых актов).

(2)
I: üks hetk.
один момент (минуточку) ОТСРОЧКА
(7.5) I: `üld`info, (.) > kinnitamata andmetel < neli kolm viis?
общая информация, по неподтвержденным данным, четыре три пять.. ИНФОРМАЦИЯ
(.)
C: jaa?
да (так?) ОБРАТНАЯ СВЯЗЬ
I: viis kaheksa,
пять восемь.. ИНФОРМАЦИЯ
C: jaa?

077-Gerassimenko_1:_Layout 1 18.05.2008 16:18 Page 497

да (так?) ОБРАТНАЯ СВЯЗЬ
I: null null.
ноль ноль. ИНФОРМАЦИЯ
C: < null null. >
ноль ноль. ОБРАТНАЯ СВЯЗЬ

3. Запросы клиентов

A. Начальный запрос

Задачей информатора является распознать цель пользователя (получить информацию) и помочь ее достичь. Обычно диалог содержит один запрос (2/3 запросов в подкорпусе – начальные), однако возможны и несколько запросов в одном диалоге.

Запрашиваемая информация разнообразна – телефонные номера (75%), адреса, приемные часы учреждений, сфера деятельности фирм. В эстонском языке запрос формулируется при помощи следующих шаблонов:

palun X `пожалуйста (будьте добры), X` или
palun öelge (mulle) X `пожалуйста, скажите (мне) X` или
ma sooviks X `я хотел(а) бы X`

где *X* обозначает пробел в знаниях клиента, который он желает заполнить.

B. Последующие запросы

К этой группе относятся 1/3 запросов: как правило, они встречаются в диалогах, в которых информатор не смог удовлетворить первичный запрос клиента (запрошенная информация отсутствует или не может быть представлена сразу) и предложил другую релевантную информацию. Такому предложению обычно предшествует уточняющий поддиалог (3).

(3)
I: .hh ma `pakuksin selle `üldise `info[`laua, jah .hh]
я бы предложила общий информационный отдел, да ОТСУТСТВИЕ ИНФ + ПРЕДЛОЖЕНИЕ
C: [jah, {oleks `hea. jah}]
да, было бы хорошо. да. СОГЛАСИЕ + ЗАПРОС
I: `number kinnitamata andmetel `seitse `kolm, `neli,
номер, по неподтвержденным данным, семь три четыре.. ИНФОРМАЦИЯ

После ответа на запрос клиент может инициировать уточняющий поддиалог, чтобы запросить добавочную информацию, относящуюся к теме запроса (4): номера телефонов, адреса, имена, сферы деятельности учреждений и т.д.

(4)
C: sooviks=seda (.) `T:oode numbrit, a- akselts `Toode.
мне нужен номер Тооде, акционерного общества Тооде ЗАПРОС
/--/
I: `kaheksa null=`null (1.0) `seitse `null,
восемь ноль ноль .. семь ноль.. ИНФОРМАЦИЯ
(0.5)
C: * mhmh *
угу ОБРАТНАЯ СВЯЗЬ
I: null=`null.
ноль ноль ИНФОРМАЦИЯ
C: null=`null.
ноль ноль ОБРАТНАЯ СВЯЗЬ
C: ja `kus ta `asub.
и где они находятся? УТОЧНЯЮЩИЙ ВОПРОС

077-Gerassimenko_1:_Layout 1 18.05.2008 16:18 Page 498

I: .hh `Tähe `sada kuus`teist,
Тяхе сто шестнадцать ИНФОРМАЦИЯ

С. Лингвистические характеристики запроса

Для автоматического распознавания диалоговых актов в нашем корпусе использовались статистические методы [7]. Точность распознавания была невысока (в среднем, около 50%), что объясняется неоднородностью корпуса, содержащего диалоги различных ситуационных типов (звонки в регистратуру, турагентства, транспортные и страховые службы), подробностью типологии, содержащей 127 диалоговых актов, и полифункциональностью речевых шаблонов. Разнообразный и подробно размеченный материал ценен для исследования естественного человеческого общения в различных ситуациях, но неудобен для тренировки статистических методов. Точность распознавания можно улучшить добавлением лингвистических правил, лексических и синтаксических ключевых сигналов, характеризующих директивные запросы.

Для формулирования запросов используется ограниченное число шаблонов. Большинство запросов включают глагол-сказуемое (12 различных глаголов): наиболее распространены глагол *paluma* (*просить*, в 1 л. ед.ч. *пожалуйста*), *ütlema* (*сказать*) и *soovima* (*желать*, табл. 2).

ГЛАГОЛ	НАКЛОНЕНИЕ (#)			ВСЕГО
	инд.	конд.	имп.	
paluma 'просить' (в 1 л. ед.ч. «пожалуйста»)	6	8		14
ütlema 'сказать'			8	8
soovima 'желать' («я хочу/ хотел бы»)	2	5		7
tahtma 'хотеть'		5		5
võtma 'брать' («дайте»)		5		5

Таблица 2. Наиболее частотные глаголы в запросах клиентов.

Глаголы-сказуемые часто используются в определенном наклонении и лице [5], как ключевые слова. Глаголы можно разделить на две группы: использующиеся в императиве (*öelge `скажите`, andke `дайте`*) и использующиеся в первом лице индикатива или кондиционалиса (*palun `прошу`* «*пожалуйста», soovin/tahan `желаю/хочу` «я хотел бы», võtaksin `я бы взял(а)` «дайте мне»*). Индикатив, универсальная форма декларативных актов, используется в нашем корпусе лишь в случае устойчивого выражения *palun* (*`прошу`* «*пожалуйста»*) для смягчения прямых директивных высказываний. Глагол *просить* может также использоваться в 1 лице кондиционалиса для усиления вежливости [8]. Признак кондиционалиса *-ks-* может служить надежным сигналом для автоматического распознавания.

Другим сигналом служит позиция глагола в реплике. Глагол начинает реплику в 22 случаях (38%), в других случаях реплика начинается с местоимения *mina/ma `я`* (16 случаев, 28%) или с дискурсивного маркера (6 случаев): *et `что` «так, то есть»`, aga `а`, äkki `вдруг` «может быть»`*. Глагол в этих случаях занимает вторую позицию.

Последующие запросы, в отличие от начальных, обычно начинаются с союза *aga `а`* или *ja `и`* или включают частицу *ka `тоже`* или *veel `еще`*. Эти слова, указывающие на продолжение, не могут быть использованы в начальных запросах (5) (6). Глагол *võtma `брать`* в кондиционалисе встречается только в последующих запросах.

(5)
C: aga `võtaks Lelula `numbri.
C: а (я) взяла бы («дайте-ка») номер Лелула

(6)
C:.hh võtaks selle `mängumaa `ka, Anni `mängumaa.
C: (я) взяла бы («дайте-ка») эту игровую комнату тоже, игровую комнату Анни

Запросы, выполняющие более одной функции (напр., согласие + запрос, см. табл. 1), нетипичны, и их форма полностью соответствует добавочной функции высказывания (прим. 3, реплика 2).

Сигналами для распознавания директивных запросов могут служить: 1) глаголы, 2) связанные с семантикой глагола морфологические формы, 3) порядок слов.

077-Gerassimenko_1:_Layout 1 18.05.2008 16:18 Page 499

4. Ответы информатора

После запроса диалог может развиваться двояко: 1) информатор выполняет запрос немедленно; 2) информатор инициирует поддиалог уточнения, проясняя запрос. Запрошенная информация предоставляется немедленно в 60% случаев (1).

В поддиалогах уточнения (40%) информатор уточняет название или расположение учреждения, номер телефона, дату запрашиваемого события и т.д. (7).

(7)
I: .hhh siis te mõtlete `Epitari.
так вы имеете в виду Эпитар? УТОЧНЯЮЩИЙ ВОПРОС
C: `j:ah, `Epitari.
да, Эпитар. УТВЕРДИТЕЛЬНЫЙ ОТВЕТ
I: jah,
да (ага) ОБРАТНАЯ СВЯЗЬ
I: üks=`hetk
один момент ОТСРОЧКА
I: (...) .hh `number on `üldinfo,=
номер - общая информация.. ИНФОРМАЦИЯ

Вторые элементы смежной пары запрос-ответ представляют собой предоставление информации, сообщение об отсутствии информации и сообщение об отсрочке (табл. 1).

Для информационного агента важно, как представить информацию наиболее естественным и в разумной степени лаконичным путем: как построить ответ, как разделить информацию на удобные для восприятия блоки.

А. Лингвистические характеристики ответов

Простейший способ сформулировать ответ – использовать в ДС готовые шаблоны возможных ответов, например:

<Учреждение> – телефонный номер <X>.

Однако информаторы в естественных диалогах имеют дело с более сложными запросами и ведут себя иначе. В 1/3 случаев используются экономичные эллиптические конструкции. Самая распространенная цель клиента – узнать телефонный номер. Информаторы разделяют телефонные номера, состоящие из 7 цифр, на две или три реплики (3, 2 и 2 цифры или 3 и 4 цифры), делая между ними паузы, в которое клиент подтверждает получение информации или переспрашивает недослышанное (2). Интонация реплик постепенно понижается, позволяя клиенту спрогнозировать конец номера; после него клиент обычно повторяет цифры последней реплики или номер целиком. В этом случае молчание информатора является подтверждением правильности понимания. Если клиент хочет эксплицитного подтверждения правильности, он оформляет повтор как вопрос: «<Номер>, да?» [9]. ДС должна учитывать, желает ли клиент подтверждения повтора.

Если ответу на запрос клиента предшествует уточняющий поддиалог, он, как правило, оформляется полным предложением: информатор маркирует возвращение к основной линии диалога, связывая реплику с запросом.

Сообщение об отсутствии информации также более подробно: информатор поясняет, почему не может предоставить информацию (8) или предлагает замещающую информацию (3).

(8)
I: ma vaatan et `selle nimega nagu ei `näita.
я смотрю, что с этим названием вроде бы не показывает.

Для сообщения об отсрочке (нерелевантной для ДС, которая оперирует базами данных во много быстрее человека) используется ограниченный список устойчивых выражений:

üks hetk `один момент` (минуточку) или
üks hetk, palun `один момент, пожалуйста` (7).

077-Gerassimenko_1:_Layout 1 18.05.2008 16:18 Page 500

В. Запросы, не получившие ответа

Первичный запрос не получает ответа в тех случаях, если он неясен информатору и в процессе его уточнения клиент формулирует новый запрос (3 случая).

5. Структура диалога

Институциональный телефонный диалог четко структурирован. В конвенциональной вводной части информатор представляется, и собеседники обмениваются приветствиями. Отвечая на звонок, информатор придерживается шаблона:
<название учреждения> <имя информатора> <приветствие>

Клиент, как правило, ограничивается ответным приветствием (9) и переходит к главному информационному запросу, открывающему главную часть диалога. Информатор может инициировать уточняющие поддиалоги после запроса, клиент – после получения ответа на запрос. Диалог заканчивается конвенциональной частью, в которой клиент благодарит информатора, вместе с тем инициируя завершение разговора, а информатор принимает благодарность и тем самым завершает разговор. Если диалог длится дольше обычного, то клиент не ограничивается благодарностью, но и эксплицитно прощается (10).

(9)
I: info`telefon=
справочная служба ИДЕНТИФИКАЦИЯ
I: Kersti=
Керсти ИДЕНТИФИКАЦИЯ
I: tere
здравствуйте ПРИВЕТСТВИЕ
C: tere päevast.
добрый день ПРИВЕТСТВИЕ

(10)
C: suur `tänu teile.=
большое спасибо вам БЛАГОДАРНОСТЬ
I: =jaa palun.
да, пожалуйста ПРИНЯТИЕ
C: head päeva.
до свидания. ПРОЩАНИЕ
I: <кладет трубку>

Для ДС важно, что 1) при выполнении запроса предпочтительны короткие реплики; 2) информация делится на короткие блоки, разделенные паузами – это позволяет клиенту записать или запомнить ее и подтвердить ее получение; 3) информатор предлагает замещающую информацию, если не может выполнить запрос клиента; 4) реплики обратной связи указывают, что информация успешно получена и партнер готов слушать дальше; 5) для приветствия, прощания, выражения согласия и обратной связи используется ограниченный набор устойчивых выражений.

Работая над ДС, мы пытаемся смоделировать «идеального» информатора – четко произносящего реплики, терпеливого и вежливого. В естественных диалогах информатор не всегда отвечает этому идеалу, однако люди-информаторы пользуются удобными стратегиями, присущими устной речи и позволяющими информатору быть лаконичным и кооперативным.

Типичная структура звонка в справочную службу представлена формальным описанием (схема 1).

диалог ::= конвенциональное_введение (запрос (уточнение)* ответ (уточнение)*)$^{+}$
конвенциональное_окончание
ответ ::= предоставление_информации | отсутствие_информации
конвенциональное_введение ::= (звонок ответ) (введение-приветствие приветствие)
конвенциональное окончание ::= благодарность принятие [прощание прощание]
уточнение ::= запрос ответ

Схема 1. Грамматика информационного диалога

077-Gerassimenko_1:_Layout 1 18.05.2008 16:18 Page 501

Грамматика представляет звонок в справочную службу как последовательность смежных пар диалоговых актов, в которых первый элемент производится одним участником диалога, а второй – его собеседником.

Смежные пары диалоговых актов предполагают использование магазинной структуры. Начальный запрос клиента задает цель, помещаемую в магазин и остающуюся там, пока запрос не будет выполнен. Каждый последующий запрос задает новую (под)цель, которая должна быть выполнена, чтобы достичь начальной цели. Цели помещаются в магазин в хронологическом, а извлекаются в обратном хронологическому порядке.

В несколько ином членении структура информационного диалога описана на материале русских справочных диалогов А.Е. Кибриком, выделяющим фазу запроса (формулирование и прием запроса), фазу нормализации запроса (промежуточные уточняющие вопросы и вспомогательные ответы) и фазу ответа (ответ оператора на запрос и снятие запроса пользователем) [11].

Для общения с пользователем на эстонском языке разработаны две простые ДС [10]. Одна из них (Транспортный агент) предоставляет информацию о вылетах из Таллиннского аэропорта. Другая ДС (Театральный агент) отвечает на вопросы о программах эстонских театров. В обе программы интегрирован синтез текст-речь. Обе системы пользуются ключевыми словами для распознавания запросов клиента и шаблонами предложений для генерирования ответов. Для ведения диалога используется регулярная грамматика.

6. Выводы

Мы проанализировали эстонские естественные справочные диалоги с дальнейшей целью разработки диалоговой системы.

Диалог состоит из трех частей: 1) конвенционального введения в диалог, содержащего идентификацию информатора и взаимные приветствия, 2) основной части, в которой информатор отвечает на запрос(ы) клиента, и 3) конвенционального завершения диалога.

Начальный запрос клиента задает цель, которая должна быть достигнута в сотрудничестве с оператором. Собеседники могут инициировать поддиалоги для уточнения запроса или ответа. Автоматическому распознаванию запросов способствуют ключевые сигналы, выявляемые в репликах запросов (определенные глагольные формы и порядок слов). Поведение информаторов в естественных диалогах экономично и лаконично, при возможности используются эллиптические конструкции. В конвенциональных фазах и репликах диалога используются устойчивые выражения. Нужно учитывать это в разработке ДС, общающейся на естественном языке. Результаты работы будут использоваться в «оестествлении» существующих диалоговых систем.

Список литературы

1. Lutz H.-D. Software-ergonomische Entwicklung – eine Herausforderung für die Computerlinguistik // SDV – Sprache und Datenverarbeitung, Vol. 25, 1, 2001, pp. 5-20.

2. McTear M.F. Spoken Dialogue Technology: Toward the Conversational User Interface. London: Springer Verlag, 2004.

3. Minker W.and Bennacef S. Speech and Human-Machine Dialogю Boston/Dordrecht/London: Kluwer Academic Publishers, 2004.

4. Hutchby I. and Wooffitt R. Conversation Analysis. Principles, Practices and Applications. Polity Press, 1998.

5. Koit M., Valdisoo M., Gerassimenko O., Hennoste T., Kasterpalu R., Rääbis A., and Strandson K. Processing of requests in Estonian institutional dialogues: corpus analysis // Text, Speech and Dialogue, Proceedings. (Eds) V. Matousek, P. Mautner, Springer, 2006, pp. 621-628.

6. Gerassimenko O., Kasterpalu R., Koit M., Rääbis A., and Strandson K. Initial requests in institutional calls: corpus study // International Conference Recent Advances in Natural Language Processing. Proceedings: Recent Advances in Natural Language Processing, (Eds) G. Angelova, K. Bontceva, R. Mitkov, N. Nicolov, N. Nikolov. Shoumen, 2007, pp. 230-234.

7. Fishel M. Complex taxonomy dialogue act recognition with a Bayesian classifier // Proc. of DECALOG Workshop on the Semantics and Pragmatics of Dialogue. Rovereto, Italy, 2007, pp. 161-162.

8. Erelt M. (Ed.) Estonian language // Linguistica Uralica Supplementary Series,vol 1. Estonian Academy Publishers, Tallinn, 2003.

9. Õim H. and Koit M. Developing a dialogue system that interacts with a user in Estonian // Inquiries into Words, Constraints, and Contexts. Festschrift for Kimmo Koskenniemi on his 60th Birthday (2005). Antti Arppe, Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna Westerlund and Anssi Yli-Jyrä (Editors). CSLI Studies in Computational Linguistics ONLINE. Copestake, Ann (Series Editor). CSLI Publications, Stanford, California. Available on-line at: http://cslipublications.stanford.edu/site/SCLO.html pp. 278-288.

077-Gerassimenko_1:_Layout 1 18.05.2008 16:18 Page 502

10. Treumuth M., Alumäe T., and Meister E. A natural language interface to a theater information database // Language Technologies, IS-LTC 2006: Proceedings of 5th Slovenian and 1st International Conference, 9th - 10 October, Ljubljana, Slovenia. (Eds) T. Erjavec, J. Žganec Gros. Ljubljana, 2006, pp. 27 - 30.

11. Кибрик А.Е. Динамика информационного диалога (на материале взаимодействия со справочно-информационной службой 09) // Очерки по общим и прикладным вопросам языкознания. Москва: УРСС, 2001. стр. 301-313.

078-Kobozeva-Sun:_Layout 1 18.05.2008 16:19 Page 503

РАСПОЗНАВАНИЕ СЕМАНТИКИ ПАДЕЖА ДЛЯ ЦЕЛЕЙ АВТОМАТИЧЕСКОГО ПЕРЕВОДА С РУССКОГО ЯЗЫКА НА КИТАЙСКИЙ: ТВОРИТЕЛЬНЫЙ ИНСТРУМЕНТА VS. ТВОРИТЕЛЬНЫЙ СРАВНЕНИЯ

RECOGNITION OF CASE SEMANTICS FOR RUSSIAN-CHINESE AUTOMATIC TRANSLATION: INSTRUMENTAL OF INSTRUMENT VS. INSTRUMENTAL OF COMPARISON

Сунь Шуан *(sunshuang@mail.ru),* ***Кобозева И.М.*** *(kobozeva@list.ru)*
Московский государственный университет им. М.В. Ломоносова

На базе «Онтологической семантики» строятся формальные правила распознавания семантических ролей инструмента и объекта сравнения (по форме и общего), кодируемых в русском языке творительным падежом, необходимые для интерпретации свободных синтаксем имен артефактов в процессе их перевода на китайский язык.

Одной из серьезных проблем при автоматическом переводе (АП) с русского языка на китайский является распознавание семантики русских падежей. Особую трудность представляет перевод падежей свободных синтаксем, то есть именных групп (ИГ), не заполняющих валентности каких-либо слов в предложении. Цель нашей работы — показать, что данная проблема может быть решена, если в системе есть семантический компонент, включающий в себя универсальную онтологию «Онтологической семантики» С. Ниренбурга и В. Раскина [5] и настроенный на нее словарь, а также правила приписывания семантических ролей. Принципиальная возможность семантической интепретации падежа, адекватной целям перевода с русского на китайский, будет продемонстрирована на примере свободных синтаксем творительного падежа без предлога (далее — ТВОР). При этом в связи с ограничениями объема доклада из всего множества возможных синтаксем ТВОР рассматриваются только те, которые могут обозначать, в зависимости от контекста, либо роль ***инструмента,*** либо роль ***объекта сравнения***. Разграничение этих значений ТВОР необходимо постольку, поскольку каждое из них переводится на китайский язык своим специальным средством. Естественно, что при формулировании правил приписывания синтаксеме этих значений мы старались учитывались и другие в принципе возможные для них значения, чтобы правильно задать условия применимости правил.

Основным средством для обозначения роли инструмента (Instr) в китайском языке является слово *yòng* (用) в синтаксической позиции предлога[1]; для выражения роли объекта сравнения (Similar) употребляется либо слово *xíng*（形） в синтаксической позиции послелога[2]; либо рамочная конструкция, состоящая из предлога *xiàng*（像）, с помощью которого выражается сходство или тождество одного предмета с другим и прилагательного *yīyàng* （一样）со значением ‘одинаковый, подобный’[3]. Выбор одного из двух способов определяется тем, какая характеристика объекта сравнения положена в основу уподобления. Если уподобление проводится по форме объекта, то ТВОР переводится при помощи *xíng*（形）, если же по иным характеристикам (например, скорости движения, характеру звучания и т.п.), то используется рамочная конструкция. Следовательно, для целей АП с русского языка на китайский целесообразно разделить роль объекта сравнения на две более конкретные роли — объекта сравнения по форме (Similar-form) и объекта общего сравнения (Similar).

Таким образом, чтобы правильно перевести свободную синтаксему ТВОР на китайский язык, в нашем случае необходимо не только установить ее общее грамматическое значение в русском языке, но и конкретизировать его. Протестировав представленные в настоящее время в Интернете системы перевода on-line с русского языка на китайский, мы убедились, что они допускают в этом деле грубые ошибки. Например, одна из систем (http://www.russky.net) предлагает для предложения (1а) перевод (1б):

---

[1] В синтаксической позиции предиката yòng (用) имеет значение ‘использовать’. Кроме этого употребляется и предлог ná (拿) с исходным предикатным значением ‘брать, поднимать’, но сфера его применения уже.

[2] В синтаксической позиции аргумента xíng（形）имеет значение ‘форма’

[3] Также для выражения сходства используется и конструкция *xiàng...sìde* (像. . . 似的), *xiàng...yībān* (像. . . 一般) и т.д, но не в тех типах контекстов, которые рассматриваются в данной работе.

078-Kobozeva-Sun:_Layout 1 18.05.2008 16:19 Page 504

(1)	а.	*Молотком*	*и*	*топором*		*выправил*	*вмятину.*
	б.	由 锤子	和	由 轴	它	调直 了	凹痕。
		Yóu chuízi	hé	yóuzhóu	tā	tiáozhíle	āohén.
		PREP молоток	и	PREP ось	он	выправил	вмятина

Мы видим, что синтаксемы ТВОР — *молотком* и *топором*, которые имеют в данном контексте значение инструмента и должны были быть переведены при помощи предлога *yòng* (用), были проинтерпретированы неверно, о чем свидетельствует выбор предлога *yóu*（由）, имеющего несколько локативно-направительных значений. Причиной данной ошибки явилась неспособность системы выбрать нужное значение ТВОР в зависимости от семантики существительного: именам, обозначающим инструменты (*молоток*, *топор*) была приписана не свойственная им локативно-направительная роль. (Одну из таких ролей ТВОР может выражать, но при именах, относящихся к семантическому классу мест и некоторым другим, ср. *лесом*, *рекой, дорогой, трубопроводом* и т.п.). В (1б) есть и другие ошибки, но они уже не имеют отношения к нашей теме. Очевидно, что, чтобы избежать ошибок при переводе свободных синтаксем ТВОР, система АП должна уметь распознавать их семантику.

Семантика свободных падежных синтаксем, то есть семантическая роль денотата имени в ситуации, обозначаемой предикатом клаузы, не определяется однозначно ни этим предикатом, ни каким-либо из его актантов, поскольку эти синтаксемы не входят в модели управления (МУ), которые должны быть сопоставлены именам и предикатам в словаре системы АП. Поэтому для установления того, какую семантическую роль выражает ТВОР в таких случаях, необходимы специальные правила, учитывающие семантические и формальные свойства как самого имени в форме ТВОР, так и слов, с которыми данная синтаксема синтаксически связана в предложении. Такие правила должны стать частью семантического компонента системы, обеспечивающего семантико-синтаксический анализ входного текста в необходимом для качественного перевода объеме. Предполагается, что семантико-синтаксическому анализу предложения предшествуют этапы морфологического и частичного поверхностно-синтаксического анализа, и начинается он с приписывания словам семантической информации, извлекаемой из словаря.

Необходимый для создания таких правил инвентарь семантических ролей, выполняемых свободными синтаксемами ТВОР, может быть извлечен из обобщения разных классификации значений ТВОР, существующих в русистике. Условия, при которых эта форма выражает те или иные роли, также описаны (см. [2]). При формулировании таких условий решающую роль играют семантические свойства лексемы, маркированной ТВОР, и семантические свойства предикатов и других слов, синтаксически связанных с данной лексемой. Однако в [2] эти свойства описаны неформально, без установки на автоматический семантический анализ. Для разработки формальных правил распознавания семантики падежей необходим семантический метаязык, включающий в себя достаточное количество четко определенных семантических единиц и заданных на них семантических отношений. Принципы построения такого языка и способы описания языковой семантики в его терминах изложены в [5]. В настоящее время построенный на основе этих принципов язык включает в себя около 6000 субстанциональных и реляционных концептов из самых разных предметных областей, в терминах которых могут быть выражены как парадигматические свойства слов (например, семантический класс, семантический признак), так и их и синтагматические семантических свойства (например, выражаемые падежами семантические роли). Концепты организованы в онтологию, представляющую собой иерархию, основанную на родо-видовых отношениях, благодаря чему семантическое правило, в котором упоминается концепт определенного уровня в онтологической иерархии, будет применимо ко всем его дочерним концептам вплоть до концептов терминального уровня. При необходимости онтология может пополняться новыми концептами, которые должны получить четкую неформальную дефиницию и формальное описание в терминах уже существующих концептов. Ниже при формулировании правил мы будем использовать именно этот метаязык.

В свете поставленной задачи наибольший интерес представляют те случаи, в которых имя в форме ТВОР (напомним, что речь идет только о свободных синтаксемах) в принципе способно выступать во всех трех вышеупомянутых семантических ролях. В русском языке существуют слова, которые могут выступать и в роли инструмента, ср. ***Колоколом*** <Instr> *жрец созывает жителей,* и в роли объекта сравнения по форме, ср. *Свод пещеры расходился* ***колоколом***<Similar-form>[4], и в роли объекта общего сравнения (то есть сравнения по каким-либо другим параметрам, напр., по звучанию), ср. *Его голос гремел* ***колоколом***<Similar>). Мы сочли целесообразным начать разработку правил распознавания значений ТВОР на материале именно таких слов. При этом мы исходили из предположения, что полученные таким образом правила можно будет экстраполировать на весь класс артефактов путем ослабления семантических ограничений на имя в форме ТВОР в правилах приписывания соответствующих ролей.

[4] Здесь и далее примеры либо взяты (как правило, в сокращенном варианте) из НКРЯ, либо получены путем трансформации нефинитных конструкций из НКРЯ в финитные предложения.

078-Kobozeva-Sun:_Layout 1 18.05.2008 16:19 Page 505

Для начала из Национального корпуса русского языка (НКРЯ) были выбраны все предложения со словом *колокол* в форме ТВОР. Их оказалось всего 209. Мы удалили из этого массива все контексты, в которых она либо употреблялась с управляющим предлогом, либо была связанной синтаксемой (ср. … *хотят* ***обеспечить*** *знаменитого звонаря-композитора* ***колоколами***), поскольку интерпретация предложных синтаксем — это отдельная задача, а интерпретация связанных синтаксем содержится в МУ слова-хозяина в словаре. После этого осталось 40 примеров на свободные синтаксемы ТВОР, которые и были подвергнуты семантическому анализу.

Устраняя из массива связанные синтаксемы ТВОР, мы обратили внимание на омонимию форм ТВОР у слова *колокол*, синтаксически связанного с глаголами звука (*греметь*, *звенеть* и т.п.)[5]. В одних случаях ТВОР выражает семантическую роль Источника звука (по [4]), как в примере (2а):

(2а) *Церковь трезвонила всеми* ***колоколами***.

Здесь ***колоколами*** — косвенное дополнение, актант предиката *трезвонить* с семантической ролью Источника звука, которая является частным случаем роли Пациенс (поскольку в общем случае объект становится источником звука не сам по себе, а в результате воздействия на него некоего Агенса-Каузатора). То, что в русском языке эта роль в предложениях с Агенсом-Каузатором кодируется формой ТВОР, по мнению Е. В. Падучевой, свидетельствует о том, что Источник звука в ситуации, когда Каузатор не совпадает с Источником звука, концептуализируется как Инструмент. В китайском же языке ИГ с ролью Источника звука в аналогичной ситуации занимает позицию прямого дополнения глагола звука, как показывает перевод примера (2а) — (2б), а значит, концептуализируется как объект, подвергающийся воздействию — Пациенс:

(2б)	教堂	敲响	了	所有的	钟。
	Jiàotáng	qiāoxiǎng-le		suǒyǒude	zhōng.
	Церковь	трезвонить-PERF		все	колокол[6]

В других случаях форма ТВОР при глаголах, которые употреблены либо в прямом значении звука (репрезентируют концепт EMIT-SOUND), либо в переносном значении ощущения (репрезентируют концепт INVOLUNTARY-PERCEPTUAL-EVENT), является обстоятельством образа действия и кодирует роль Объекта сравнения Similar, как в (3), где глагол *гудеть* выступает в переносном значении:

(3) *Голова гудела колоколом*.

В развернутом виде смысл (3) можно представить в виде перифразы ‘В голове говорящего локализовалось ощущение, подобное ощущению звука гудящего колокола’. Соответственно при переводе (3) на китайский язык используется конструкция *xiàng ... yīyàng*（像. . . 一样）:

(3’)	头	像	钟	一样	响。
	Tóu	xiàng	zhōng	yīyàng	xiǎng.
	голова	подобно	колокол	одинаковый	гудеть

Для разрешения подобной синтаксической (а часто также и сопровождающей ее лексической) омонимии необходимо, чтобы в словаре системы АП, во-первых, при ЛСВ существительных были указаны соответствующие им онтологические концепты, они же семантические классы (например, *церковь*-1 — CHURCH, *церковь*-2 — CHURCH-BUILDING), а во-вторых, глаголы были представлены всеми возможными для них МУ с указанием семантических ограниченииий по валентностям. Тогда при анализе (2) для глагола звука *трезвонить* будет выбрана МУ с диатезой типа процесс-каузация (4):

(4) *трезвонить*-1 EMIT-SOUND

МУ 1

Синт. актант	1 — ИГ им	2 — (ИГ тв)
Сем. роль	1 — Agent	2 — Patient-Source
Сем. ограничение	1 — [HUMAN, ORGANIZANION]	2 — [DEVICE, MUSICAL- INSTRUMENT]

Действительно, из двух ЛСВ вокабулы *церковь* в контексте данного глагола может быть выбран только вариант *церковь*-1, который относится к семантическому классу CHURCH (/RELIGIOUS-ORGANIZATION/ORGANIZATION)[7], а *колокол* через сопоставленный ему концепт BELL входит в классы

[5] Естественно, что такая омонимия свойственна не только данному слову, но всем словам, относящимся к классу артефактов, предназначенных для извлечения звуков — *труба, бубен, гудок* и т. п., если они способны выражать роль Источника звука в ситуации звучания формой ТВОР (ср. *бренчать бубенчиком*, но *бренчать на гитаре*).

[6] Такой же перевод будет иметь и предложение *Церковь трезвонила во все колокола*, где роль Источника звука кодируется конструкцией «*в* + вин. пад.», выбор которой отражает концептуализацию этого участника ситуации как Конечной точки удара, при помощи которого Каузатор вызывает его звучание, ср. *бить в барабан, стучать в окно* и т.п.

[7] Через знак «/» мы показываем путь, связывающий данный концепт с иерархически вышележащими, родовыми для него концептами.

078-Kobozeva-Sun:_Layout 1 18.05.2008 16:19 Page 506

устройств — BELL/DEVICE и музыкальных инструментов — BELL/MUSICAL-INSTRUMENT Соответственно синтаксема ТВОР будет распознана уже на этом шаге анализа как имеющая роль Patient. Та же МУ представлена в примере (5):

(5) *Я* ***трезвонил шпорой***.

При анализе (3) для глагола *гудеть* будет выбрана МУ (6)[8]:

(6) *гудеть*-n INVOLUNTARY-PERCEPTUAL-EVENT

Синт. актант	1 — ИГ им	2 — (от ИГ род)
Сем. роль	1 — Location	2 — Cause
Сем. ограничение	1 — ANIMAL-PART (not (MOUTH, LARYNX))[9]	2 — [EVENT]

Поскольку значение синтаксемы ТВОР на этом шаге анализа не будет распознано, она поступит на обработку в блок правил для анализа свободных синтаксем, где ей должна быть приписана роль объекта сравнения Similar.

Ниже мы представим те правила, которые были первоначально сформулированы на базе 40 примеров со словом *колокол*, а затем проверены на материале еще 680 примеров из НКРЯ с существительными ТВОР из класса артефактов. Правила имеют следующий вид. В левой части правила задается исходный шаблон, содержащей синтаксему ТВОР с незаполненной семантической ролью — <?>, а в правой части правила приводится та же синтаксема с заполненной ролью. Поясним принятые нами нотационные конвенции на примере одного из правил приписывания роли Similar-form, обозначаемого ярлыком R1-Similar-form:

- **R1-Similar-form**: V: CHANGE-POSITION + $[_{NP}\ [N_{ins,sg}$: GEOMETRICAL-OBJECT, PHYSICAL-OBJECT{FORM-TEMPLATE}$]]_{NP}$ <?> → $NP_{ins,sg}$<Similar-form>

При записи правила помимо обычных конвенций, принятых в грамматике НС, при именных и глагольных синтаксемах указывается вся необходимая информация. Во-первых, это морфологическая информация в виде ярлыков, используемых для этой цели в НКРЯ, записываемая при лексической категории как ее нижний индекс (отсутствие морфологического ярлыка означает, что соответствующий морфологический признак не играет роли). В **R1-Similar-form** это творительный падеж (ins) и единственное число (sg). Во-вторых, это семантический класс лексической категории, указываемый при ней после двоеточия и обозначаемый именем онтологического концепта (если имя может относиться к нескольким семантическим классам, то они перечисляются через запятую). В **R1-Similar-form** для N это классы геометрических объектов — GEOMETRICAL-OBJECT (*круг*, *крест* и т. п.) и физических объектов — PHYSICAL-OBJECT, а для V — класс пребывания в позиции (*сидеть*, *торчать* и т.п.) или изменения позиции (*садиться*, *вытягиваться*, *поворачивать* и т. п.) — CHANGE-POSITION[10]. При имени семантического класса в фигурных скобках может быть указан нетривиальный семантический признак слова [1]. В **R1-Similar-form** это признак «эталон формы» — FORM-TEMPLATE, свойственный группе слов из разных классов онтологии, напр., словам *колокол* (класс ARTIFACT), *груша* (FOODSTAFF), *змейка* (ANIMAL) и многим др. [3]. Нетривиальные семантические признаки — еще один вид семантической информации, приписываемый слову в словаре системы. Слово *колокол* имеет как минимум два нетрививальных семантических признака: «эталон формы» и «эталон звука» — SOUND-TEMPLATE. Наконец, после парадигматических семантических признаков имени указывается его семантическая роль в угловых скобках, обозначаемая при помощи имен из общепринятого инвентаря семантических отношений с некоторыми необходимыми добавлениями — в **R1-Similar-form** это роль объекта сравнения по форме — Similar-form. Правило позволяет приписать свободной синтаксеме ТВОР, принадлежащей к семантическому классу геометрических объектов или физических объектов, помеченных в словаре признаком FORM-TEMPLATE, роль Similar-form, если эта синтаксема непосредственно связана с глаголом семантического класса CHANGE-POSITION.

По этому правилу устанавливается значение граммемы ТВОР (ins) в 11 случаях из нашей выборки, например:

(7) а) *Плащ* ***ниспадал колоколом*** *почти до земли.* б) *Ветерок забрался по ее ногам под юбку и* ***надул*** *платьице* ***колоколом***. в) *Кургузая шинель* ***топорщилась колоколом***. г) *Лазурь* ***колоколом стояла*** *над его тихо работаю-*

[8] Та же МУ представлена в примерах ***Гудели ноги***; …удовлетворенно ***гудят руки***… и т.п.

[9] Данное семантическое ограничение исключит выбор данной МУ (и соответствующего ей ЛСВ *гудеть*) при синтактико-семантическом анализе примеров типа *Все эти отверстые* ***рты*** *звенят и* ***гудят****, а я мысленно затыкаю их огромными кусками хлеба,* где *гудеть* репрезентирует концепт EMIT-SOUND и имеет одновалентную модель с первым и единственным актантом в роли Пациенса-Источника звука.

[10] В «Онтологической семантике» значения глаголов изменения позиции и соответствующих им глаголов пребывания в позиции (повернуть – повернуться, сажать-сидеть, вставать – стоять и т. п.) относятся к одному онтологическому концепту CHANGE-POSITION (во всех его видах).

078-Kobozeva-Sun:_Layout 1 18.05.2008 16:19 Page 507

щим станком. д) *Там стены уже напрочь* ***расходились колоколом.***

Следующее правило приписывает ТВОР то же самое значение Similar-form, когда эта синтаксема распространяет ИГ, выступая комплементом имени, относящегося к семантическому классу «одежда» (CLOTHING-ARTIFACT):

- **R2-Similar-form** $[_{NP}$ N^1: CLOTHING-ARTIFACT $[_{NP}$ $N^2_{ins,sg}$: GEOMETRICAL-OBJECT, PHYSICAL-OBJECT{FORM-TEMPLATE}$]_{NP}$ $]_{NP}$ <?> → $NP_{ins,sg}$ < Similar-form>

По этому правилу устанавливается значение Similar-form для граммемы ТВОР в 8 случаях из нашей выборки, например:

(8) а) *Вот он позирует перед кинокамерой в отражающем солнце белом* ***мундире колоколом***. б) ***Юбка колоколом*** *выше колен.* в) *Народ ... сменил старые шубейки на* ***пальто колоколом***. г) *Маленькая женщина в* ***шубке колоколом*** *… радостно, но не громко вскричала …*

Следующии три правила приписывают роль объекта общего сравнения — Similar (в нашей выборке представлена 12 случаями):

- **R1-Similar** ($[_{NP}$ N^1_{nom}: NOT ANIMAL$]_{NP}$) + V: EMIT-SOUND + + $_{NP}[N^2_{ins}$: PHYSICAL-OBJECT{SOUND-TEMPLATE}$]_{NP}$ <?> → → NP_{ins} <Similar>
- **R2-Similar** ($[_{NP}$ N^1_{nom}:HUMAN$]_{NP}$) + (V:SPEECH-ACT) + $[_{NP}$ N^2_{ins}: PHYSICAL-OBJECT{SOUND-TEMPLATE}$]_{NP}$ <?> + "DIRECT-SPEECH" → NP_{ins} <Similar>
- **R3-Similar** $[_{NP}$ N^1_{nom}:ANIMAL-PART, ANIMAL-SYMPTOM$]_{NP}$ + V: INVOLUNTARY-PERCEPTUAL-EVENT + $[_{NP}$ N^2_{ins}: PHYSICAL-OBJECT{SOUND-TEMPLATE}$]_{NP}$ <?> → NP_{ins} < Similar>

По правилу **R1-Similar** семантика ТВОР распознается в примерах типа (9):

(9) а) *… из Кремля* ***разнеслось*** *набатным* ***колоколом***: «К вам обращаюсь я, друзья мои!». б) ***Имя****-то какое! - так и* ***гудит колоколами*** *древних русских сторожевых монастырей!* в) *И так стукнул в сердцах дверью, что весь* ***дом загудел колоколом*** *- бом!* г) *Она подталкивала каталку, и* ***половник колоколом гремел*** *в кастрюле.* д) *Но самое главное, многие иерархи не готовы к реформе церкви, хотя* ***нужда*** *в ней* ***колоколами гремит*** *над землей России.* е) … в его голове ***колоколом ударил***[11] ***голос*** *Кавабаты…*

Правило **R2-Similar** рапознает семантику ТВОР в примерах типа (10):

(10) а) - ***Вспомним всех поименно! - колоколом загудел князь***. б) *…****извозчик*** *наклонил лохматую голову и* ***прогудел - колоколом*** *из подземелья: -* ***Премного благодарны***. в) И тут ***колоколом*** в голове: **«Шпионы»**.

R3-Similar служит для интерпретации ТВОР в примерах типа (11):

(11) а. ***Голова гудела колоколом***...; б) … *у меня таки* ***башка колоколом звенит.***

Напомним, что эти правила применяется уже после того, как с помощью МУ глаголов установлены семантические роли всех их актантов (связанных синтаксем), в том числе и тех, которые маркируются творительным падежом. Поэтому правило **R1-Similar** не приведет к ошибке в контекстах типа (2а), где NP_{ins} выступает в роли Пациенса-Источника звука. При этом приходится признать, что для некоторых глаголов невозможно в локальном контексте клаузы однозначно установить значение ТВОР. Так, в предложениях типа (12) без дополнительных знаний об описываемой ситуации невозможно снять омонимию Patient-Source / Similar:

(12) *Шторм тревожно* ***гудел*** *огромным* ***колоколом***.

С одной стороны, здесь применима МУ *гудеть 1*

Синт. актант	1 — ИГ им	2 — (ИГ тв)
Сем. роль	1 — Agent	2 — Patient-Source
Сем. ограничение	1 — CURRENT, WEATHER	2 — INANIMATE

выбор которой обусловливает интерпретацию колокола как реального объекта, на который воздействует шторм, вызывая его звучание. Ср. однозначный пример с той же МУ (13):

(13) ***Ветер гудит*** *рваным листовым* ***железом*** <Patient-Source>

С другой стороны, может быть выбрана и МУ *гудеть* 2

Синт. актант	1 — ИГ им
Сем. роль	1 — Agent-Source
Сем. ограничение	1 — [CURRENT, WEATHER, INANIMATE]

и тогда NP_{ins} из примера (16) поступит в блок обработки свободных синтаксем, где будет проинтерпретирована как объект сравнения (Similar) по правилу **R1-Similar**. Заметим, что даже достаточно широкий контекст, в кото-

[11] В контексте N_{nom} классов MUSICAL- INSTRUMENT, TIMEPIECE глаголы, в своем прямом значении относящиеся к классам HIT или BEAT, выступают в переносном значении EMIT-SOUND., ср. *Барабаны ударили с двух сторон; Бьют часы на Спасской башне*. Можно признать такой же семантический переход и в контексте N_{nom} класса HUMAN-VOICE, как в данном примере.

078-Kobozeva-Sun:_Layout 1 18.05.2008 16:19 Page 508

ром (16) приводится в НКРЯ не позволяет понять, какая из интерпретаций имелась в виду автором высказывания.

Ясно, что 3 рассмотренные правила для роли Similar имеют ограниченную сферу применения. Они приписывают эту роль синтаксеме ТВОР только в контексте глаголов звука, речи и неконтролируемого восприятия и только если лексема в форме ТВОР помечена в словаре признаком {SOUND-TEMPLATE}. Для приписывания той же роли в контексте других глаголов понадобятся аналогичные правила, в которых будут фигурировать глаголы других онтологических классов, а имя в ТВОР будет помечено в словаре другим нетривиальным признаком из той же «эталонной» серии (X-TEMPLATE), например, правило R4-Similar:

- **R4-Similar**: (NP^1_{nom}) + V: MOTION-EVENT + $_{NP}[N^2_{ins}$: PHYSICAL-OBJECT{SPEED-TEMPLATE}$]_{NP}$ <?> → NP_{ins} <SIMILAR>

для интерпретации ТВОР в примерах типа (14):

(14) а) … *она…**стрелою выскочила** из комнаты*; б) Мяч ***пулей влетел** в ворота*; в) ***Молнией кинулась** откуда-то взявшаяся бабочка*.

Наконец, роль Instr приписывается свободной синтаксеме ТВОР по правилу R-Instr:

- **R-Instr**: ($[_{NP}$ N^1_{nom}:HUMAN, ORGANIZATION, DEVICE$]_{NP}$) + V + NP^2 <PATIENT, THEME> + [NP [N^3_{ins}]]<?> → NP^3 <Instr>

Это правило должно применяться после правил обработки свободных синтаксем ТВОР, рассмотренных выше, в соответствии с общим принципом применения правил в порядке от более частных к более общим, поскольку оно, в отличие от правил приписывания Similar и Similar-form, не ограничивает ни семантический класс имени в ТВОР, ни семантический класс глагола, требуя только присутствия при нем ИГ с ролью Пациенса или Темы (то есть, фактически, его переходности). По этому правилу роль Instr будет приписана свободной синтаксеме ТВОР в примерах типа (15), которых в нашей выборке встретилось 7:

(15) а) … *она **разбудит меня колоколом** громкого боя* …б) ***Колоколом** жрец **созывает жителей.*** в) *А я вам, сударыня, говорю, что вы **меня** этими **колоколами выгоните** из дому.* г) ***Пробили** водяную **тревогу** - протяжными гудками и **колоколом**.*

Полученные на материале слова *колокол* правила распознавания ТВОР были затем проверены на других словах класса артефактов, способных иметь все три рассматриваемые семантических роли или хотя бы две из них. Выбрав 200 первых документов со словом *труба* в ТВОР, мы убедились, что для него во всех его конкретных значениях действуют уже рассмотренные правила распознавания, но только, как и следовало ожидать, семантические ограничения на N^1 в правиле R2-Similar-form потребовалось ослабить, допустив отнесение ядерного имени к классу частей тела. В уточненном виде:

- **R2'-Similar-form:** $[_{NP}$ N^1: ANIMAL-PART, CLOTHING-ARTIFACT $[_{NP}$ N^2_{ins}: GEOMETRICAL-OBJECT, PHYSICAL-OBJECT{FORM-TEMPLATE}$]_{NP}$ $]_{NP}$ <?> → NP_{ins} <SIMILAR-FORM>

это правило сможет приписать синтаксеме ТВОР роль объекта сравнения уже не только в случаях типа (8), но и в случаях типа (16):

(16) а) ***Нос трубой**, телеса крепенькие, как брюква, мозги жидкие, как вокзальные щи.* б) *Без седла, китель расстегнулся,..., а у лошади **хвост трубой**.*

а также во многих других случаях с эталонами формы из класса артефактов или других семантических классов, ср. *губки бантиком*, *грудь колесом, нос грушей* и т.п.

Анализ первых 100 примеров со словом *серп* ТВОР, подтвердив эффективность уже введенных правил, потребовал ввести еще одно правило приписывания роли Similar:

- **R5-Similar** $[_{NP}$ N^1_{nom}: NOT ANIMAL$]_{NP}$ + V: EMIT-LIGHT + $_{NP}[N^2_{ins,sg}$: CUTTING-IMPLEMENT$]_{NP}$ <?> → NP_{ins} <SIMILAR>

для случаев, когда *серп* или другие слова класса «режущих инструментов» выступают в роли объекта сравнения (по признаку «свечения»), как в (17):

(17) а) *Вдали лунным **серпом светилась** река*; б) *На меня **блеснул серпом** ее взгляд, полный молодых чувств, дружбы, юмора.*; в) ***Блеснула** сабельным **лезвием** полоска реки.*

Также только одно дополнительное правило группы Similar дали первые 100 примеров со словом *мешок*:

- **R6-Similar** NP^1_{nom} + V: CHANGE-POSITION, MOTION-EVENT + $_{NP}[N^2_{ins}$: BAG$]_{NP}$ <?> → NP_{ins} <SIMILAR>

ср. примеры типа (18):

(18) а) *Митя **мешком сел** на пол; б) Он **мешком болтался** на турнике и не мог выполнить самого простого упражнения; в) … парашюты **мешком падали** вместе с ними. г) Форма **сидела** на нем классическим **мешком**. д) Это сейчас прошлое **висит** огромным **мешком** за плечами...*

А просмотр всех 72 примеров со словом *куль* ТВОР, подтвердив правило R6-Similar, показал, что необходимо и правило приписывания роли Similar ИГ, распространяющей имя, аналогичное правилу R2'-Similar-form:

078-Kobozeva-Sun:_Layout 1 18.05.2008 16:19 Page 509

- **R7-Similar** [$_{NP}$ N^{1}: ANIMAL-PART, CLOTHING-ARTIFACT [$_{NP}$ N$^{2}_{ins}$: BAG]$_{NP}$] $_{NP}$ <?> → NP$_{ins}$ <SIMILAR>
  ср. пример (19):

(19) ... *глазки косые,* ***спины кулем****, носы блестят*...

Интересно, что именно через призму китайского языка, в котором ТВОР в примерах типа (18) и (19) не может переводиться при помощи послелога *xíng* (形) ('форма'), видно, что слова *мешок* и *куль* не являются настоящими эталонами формы, хотя по контекстам употребления и близки к таковым. Причина понятна: обозначаемые ими объекты входят в онтологический класс мягких контейнеров (BAG), который характеризуется именно отсутствием постоянной формы.

Анализ примеров с формой ТВОР слов *молоточек* (всех 107), и *лопата* (первых 100), подтвердил применимость рассмотренных выше правил для распознавания семантики свободных синтаксем, не потребовав ни их модификации, ни введения каких-либо дополнительных правил.

Ясно, что не проанализировав аналогичным образом рассмотренные 3 значения свободных синтаксем ТВОР, представленных словами других онтологических типов (животных, растений, светил и т. п.), а также другие значения свободных синтаксем ТВОР, которые встречаются у имен артефактов, мы, конечно, не можем гарантировать, что правила групп R-Similar, R-Similar-form и R-Instr в их настоящем виде обеспечат правильное распознавание соответствующих ролей для всех имен артефактов. Однако не вызывает сомнений, что предлагаемый онтологической семантикой метаязык позволяет без особого труда представить в формальном виде всю семантическую информацию, которая может потребоваться для того, чтобы задать точные ограничения в правилах семантической интерпретации падежных граммем.

После того, как при помощи этих правил значения семантических ролей ИГ в клаузе определено, они могут быть переведены на китайский язык при помощи правил, задающих переводные соответствия семантических ролей:

NP <Instr> = yòng （用） + NP

NP <Similar-form> = NP + xíng（形）

NP <Similar> = xiàng (像) + NP + yīyàng（一样）

Так, при семантическом анализе предложения (3) правило R3-Similar припишет синтаксеме ТВОР роль <Similar>, и в его переводе (3') будет использована рамочная конструкция *xiàng* （像）+ NP + *yīyàng*（一样）. В предложении (20а) по правилу R2'-Similar-form синтаксеме ТВОР будет приписана роль <Similar-form> и оно с учетом правил порядка слов будет переведено как (20б):

(20) а. *Девушки носят голубые юбки* ***колоколом****.*

б.	姑娘们	穿着	蓝色的	钟	形	裙。
	Gūniángmen	chuāngzhe	lánsède	**zhōng**	**xíng**	qún.
	Девушки	носят	голубой	колокол	форма	юбка.

В предложении (21а) синтаксеме ТВОР будет приписана та же роль, но уже по правилу (R1-Similar-form):

(21) а. *Шинель топорщилась* ***колоколом****.*

б.	外套	扎煞成了	钟	形。
	Wàitào	zhāshàchéngle	**zhōng**	**xíng**
	шинель	топорщилась	колокол	форма.

А в предложении (22а), в котором нет условий для приписывания синтаксеме ТВОР ролей объекта сравнения (равно как и ролей маршрута, количества и др., которые мы здесь не рассматривали), ей правилом R-Instr будет приписана роль Instr и в переводе использован предлог *yòng* (用):

(22) а. ***Колоколом*** *жрец созывает жителей.*

б.	祭司	用	钟	召集	居民。
	Jìsī	yòng	zhōng	zhāojí	jūmín.
	жрец	Instr	колокол	созывать	житель

Заключение

Повышение качества перевода с русского языка на китайский предполагает решение задачи распознавания семантических ролей ИГ, выражаемых в русском языке падежными показателями. Для решения этой задачи система АП должна иметь словарь, в котором русской лексеме приписаны не только ее лексико-грамматическая категория, МУ с ограничениями по валентностям и переводной эквивалент, но и семантическая информация — онтологический класс и нетривиальные семантические признаки, а также блок правил распознавания семантических ролей свободных синтаксем, формулируемых в терминах морфологических, синтаксических и семантических единиц и отношений. В качестве языка формальной репрезентации семантической информации может быть использован язык «Онтологической семантики» С. Ниренбурга и В. Раскина, при необходимости расширяемый за счет введения дополнительных субстанциональных или реляционных концептов или уточнения существующих.

078-Kobozeva-Sun:_Layout 1 18.05.2008 16:19 Page 510

Список литературы

1. Апресян Ю. Д. Типы информации для поверхностно-семантического компонента модели «Смысл <=>Текст» // Апресян Ю.Д. Избранные труды. М., 1995. Том II.

2. Золотова Г.А. Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса. М.: Эдиториал УРСС, 2001.

3. Кобозева И. М. Как мы описываем пространство, которое видим: форма объектов. // А. С. Нариньяни (ред.), Труды международного семинара «Диалог 2000» по компьютерной лингвистике и его приложениям. Т.1. Протвино, 2000.

4. Падучева Е.В. О парадигме регулярной многозначности (на примере глаголов звука) // НТИ. Серия 2. 1998. № 4.

5. Nirenburg S., Raskin V. Ontological Semantics. Cambridge, MA: MIT Press, 2004.

079-Suhova_N:_Layout 1 18.05.2008 16:21 Page 511

СПОСОБЫ ВЗАИМОДЕЙСТВИЯ ПАУЗ КОЛЕБАНИЯ И КИНЕТИЧЕСКИХ ФРАЗ*
THE DIRECTIONS OF INTERACTION BETWEEN HESITATION PAUSES AND KINETIC PHRASES

Сухова Н.В. *(sukhova@spa.msu.ru),*
Московский государственный университет им. М.В. Ломоносова

Статья посвящена выявлению возможных направлений изучения способов взаимодействия пауз колебания и кинетических фраз в английской спонтанной речи. На материале одного звукового отрывка применяется междисциплинарный подход к изучению фонетических и кинетических средств. Результат представлен семью способами совместного исследования пауз и кинетических фраз.
Ключевые слова: пауза колебания, кинетический жест (фраза), акустические характеристики.

1. Актуальность изучения пауз в рамках невербальной семиотики

Изучение пауз началось в 50-х годах XX века и проводилось в рамках нескольких наук, таких как лингвистика (фонетика, текстология, дискурсивный анализ, переводоведение, лексикология) и психолингвистика. Анализ работ показал, что основным материалом исследований были только аудиозаписи, которые: 1) детально рассматривались с позиций просодических характеристик тех участков речи, где была зафиксирована остановка звучания, и участков, прилежащих к ним (B. Zellner 1994 и см. некоторые работы в Proceedings of the 15th International Congress of Phonetic Studies 2003); 2) служили материалом для исследований места и в дальнейшем роли пауз в потоке речи, в процессе порождения и восприятия речи и т.п. (частично О.А. Александрова 2004, M. Cecot 2001, И.Н. Рыбка 2007 и др.); 3) способствовали выявлению сопутствующих паузам или заполняющих их феноменов, таких как «слова-паразиты» (см. Ю.В. Дараган 2000, 2003) или вокализации и т.д. (см. L.Cerrato, M.D'Imperio 2003).

В рамках науки невербальная семиотика паузы практически не изучались, т. е. паузы не рассматривались совместно, например, с кинетическими жестами. Однако в связи с терминологической неточностью встречаются упоминания о невербальных видах хезитационных пауз (см. О.А. Александрова 2004), в которые включаются незаполненные и паралингвистические паузы колебания (покашливания, вздох, смех, прочищение горла, цоканье языком и кинетические элементы).

Активная разработка тем невербальной семиотики может способствовать тому, что и паузы, которые очень во многом соприкасаются с жестами, в широком смысле слова, будут исследоваться с точки зрения новых, ранее неизвестных аспектов.

Так, речь здесь может идти о: а) возможности изучения взаимодействия пауз и кинетических жестов, совместно функционирующих в речи; б) формальных, семантических, синтаксических и прагматических характеристиках взаимодействующих жестов и пауз; в) о возможных паттернах; г) о когнитивных механизмах появления таких паттернов и т.д.

При близком рассмотрении оказывается, что на данном этапе созрела необходимость изучения пауз и невербальных средств (или, более узко, кинетических жестов), совместно функционирующих в речи. Ценность такого междисциплинарного и интегрального подхода заключается в том, что система невербальной семиотики пополнится новыми элементами (возможно комплексами пауз и кинетических фраз), а просодический элемент – пауза – будет описан потенциально не только в терминах фонетики.

Цель данной работы – провести пилотный эксперимент по выявлению возможных способов взаимодействия кинетических фраз и пауз (пауз колебания) в английской спонтанной речи[1].

Материалом исследования служит отрывок из телевизионной программы ***Jamie's School Dinners*** (Channel 4: http://www.channel4.com/life/microsites/J/jamies_school_dinners/). Звуковой файл (Jamie3) был создан и обработан с помощью программы *SoundEdit Pro v. 2.1.126* и проанализирован с помощью *Praat v. 5.0.08*.

* Работа выполнена при финансовой поддержке РГНФ в рамках научно-исследовательского проекта № 08-04-00165а.

[1] В силу того, что это пилотный эксперимент, мы выносим за скобки множество релевантных лингвистических и экстралингвистических факторов, таких как акцент, пол, социальный статус, возраст и др.

079-Suhova_N:_Layout 1 18.05.2008 16:21 Page 512

2. Виды и характеристики пауз

Существуют разные **виды пауз**. Например, Г.И. Бубнова (1999: 4) классифицирует паузы на респираторные (дыхательные) и речевые (намеренные, смысловые), которые в свою очередь подразделяются на синтаксические (разделительные и соединительные) и свободные (стилистические, дискурсивные и прагматические, т.е. фатические, аппелятивные и хезитационные). Б. Зеллнер (1994) опирается на физическую / лингвистическую и психолингвистическую классификации пауз. В основе такого деления лежит противопоставление, с одной стороны, паузы как акустического явления (перерыва фонации), а с другой – как перцептивного (восприятие заполненных и незаполненных пауз).

Мы будем рассматривать паузы колебания (хезитации).

Пауза хезитации с **формальной** точки зрения представляет собой фонационный предпаузальный участок в виде хезитации и физическую паузу. В устной речи это одна из важных структурирующих категорий на интонационном уровне. В виду того, что существуют паузы, регистрируемые а) только акустической аппаратурой, б) только аудиторами, в) и аппаратурой и людьми, в данном исследовании проводится анализ пауз третьего вида, т. е. тех, что фиксируются и фонетическими программами и аудитором (автором работы).

Основными **формальными способами** реализации пауз колебания являются перерыв в фонации (адекватное восприятие паузы возникает при ее длительности от 150-200 мс и выше)[2] и фонетические и лексико-синтаксические средства.

Итак, предполагается, что фонетические и лексико-синтаксические средства ЗАПОЛНЯЮТ паузы колебания. Разные исследователи по-разному подходят к проблеме классификации подобных заполнителей (ср.: Г.И. Бубнова 1994, О.А. Александрова 2004). Если свести их мнения воедино, то получим следующие компоненты:

1) **удлинение звуков**, входящих в сегментную последовательность речевого сигнала («*по телеви-изору нра-авятся э-эти познавательные передачи*«; Александрова 2003: 96);
2) **звуковые заполнители, вокализации** [*э-э, м-м, гм*];
3) **слова-паразиты, пустые слова** (*тот самый, значит, как бы* и т.д.);
4) **метатекстовые комментарии, вводные конструкции** (*видите ли, как Вы знаете* и пр.);
5) **непреднамеренные повторы** (полный или частичный повтор слова, повтор словосочетания, повтор конструкции; например: А: «*С какой целью вы обычно смотрите телевизор?»* Б: *«С целью обычно новости узнать*«; Александрова 2003: 96);
6) **плеоназмы** (*очень прекрасный*);
7) **стереотипы, штампы** (*Ради Бога*);
8) **невербальные заполнители** (*покашливания, цоканье языком* и т.д.)[3];
9) **самопрерванные конструкции** (*фальстарты, рестарты,* разного рода *сбои* в высказывании; например, «то есть не для пу.. не предлагается для обсуждения»; Александрова 2003: 96).

Кроме упомянутых элементов, предположительно можно включить в этот список и невербальные средства, а более конкретно, **кинетические фразы**, которые мы собираемся рассмотреть на участке пауз колебания.

3. Кинетические фразы

Синтагма сопровождается жестовой фразой, которая состоит из комплекса кинетических жестов, и функционирует как единое целое. Мы рассматриваем жесты рук, движения головы, корпуса (телодвижения и позы) и выражения лица. Жесты сами по себе обладают **формальными характеристиками**[4], которые можно сгруппировать так: амплитуда, сила воспроизведения, скорость и направление (см. табл. 1).

Так, отдельно жесты характеризуются рядом параметров, которые варьируют в зависимости от степени их интенсивности. Изменяясь во времени и в пространстве, рабочий активный орган жеста достигает некой максимальной точки интенсивности, при условии, что изначально он находится в состоянии покоя. При этом необходимо учитывать, что в процессе производства *жест любой интенсивности* проходит фазу экскурсии, фазу воспроизведения и фазу рекурсии. Схематично реализацию жеста можно показать следующим образом:

1) интенсивность:

состояние покоя → жест слабой интенсивности → жест средней интенсивности → жест сильной интенсивности → **жест максимальной интенсивности;**

[2] Перцептивная оценка пауз одинаковой длительности варьирует в зависимости от синтаксического и семантического окружения, от темпа произнесения, от просодического оформления предпаузального сегмента и пр.

[3] Как уже говорилось выше, здесь наблюдается терминологическая неточность, которая будет снята ниже. Здесь классификация дается по О.А. Александровой (2003) и Г.И. Бубновой (1994).

[4] Рассматриваются полюса характеристик, но подразумевается, что у них есть и средние (нейтральные) значения.

079-Suhova_N:_Layout 1 18.05.2008 16:21 Page 513

2) фазы:

экскурсия → пик→ рекурсия.

Так, все вышеуказанные жестовые характеристики имеют две экстремальные точки: исходную и максимальную.

Характеристики жестов	***Формы кинетического поведения***			
	Руки (Р)[5]	**Голова (Г)**	**Выражение лица (ВЛ)[6]**	**Корпус тела: телодвижения (ТД), позы (П)**
Амплитуда	большая / малая (б / м)	большая / малая	большая / малая	большая / малая
Сила воспроизведения	сильный / слабый (с / сл)	сильный / слабый	сильный / слабый	сильный / слабый
Скорость	быстрый / медленный (бр / мд)	быстрый / медленный	быстрый / медленный	быстрый / медленный
Направление	вперед (от себя), назад (к себе), влево, вправо, вверх, вниз	вперед, назад, влево, вправо, вверх, вниз	(взгляд) вперед, назад, влево, вправо, вверх, вниз	вперед, назад, влево, вправо, вверх, вниз

Таблица 1. Характеристики форм кинетического поведения (подробнее см. Н.В. Сухова 2006)

В составе жестовой фразы ее компоненты, т. е. кинетические жесты любой формы и интенсивности, могут быть кинетически выделенными. Под кинетической выделенностью понимается значимое и видимое изменение положения и характеристик активных рабочих органов (Г.Е. Крейдлин 2001) жеста на звучащем отрезке. Функция *активных* органов состоит в задании способа реализации жеста. Существуют еще и *пассивные* органы жеста, которые определяют место его реализации. Выделенность жестов фиксируется по релевантным характеристикам и в соответствии с положением активного рабочего органа жеста на шкале интенсивности (ближе к точке покоя или к максимуму).

4. Способы взаимодействия пауз колебания и кинетических фраз

Рассмотрим пример **Jamie_1**[7].

I've decided |[8] I am gonna devote the next year of my life at least (**1**) ↓|[9] to try and make some big change (**2**) ↓in (**3**) ↓not just (**4**) |***er***[10] – one school (**5a**) ↓***but*** (**5b**) |***you know*** my dream is to take over a (**6**) ↓borough | or at least a huge part of a borough |…

В данном эпизоде помимо синтаксических (межсинтагменных) пауз наблюдаем и участки пауз колебания (6 случаев).

Паузы входят в состав просодического контура наряду с другими звучащими элементами. В силу этого ее **формальными** акустическими признаками являются длительность (duration), высота основного тона (pitch) и интенсивность (intensity).

Общая длительность интересующего нас отрывка 14,27 с.

Основной мелодический контур этого участка выглядит так: см. Рис. 1.

Кривая интенсивности представлена следующим образом (Рис. 2).

[5] Условные обозначения, принятые в работе.

[6] Имеются в виду движения, связанные с глазами, бровями, щеками, носом, подбородком и губами (челюстями).

[7] Название файла. Говорит Джейми Оливер, организатор проекта и главный участник проекта **Jamie's school dinners** (см. http://www.jamieoliver.com/).

[8] Межсинтагменная пауза.

[9] Пауза колебания.

[10] Заполнители паузы колебания.

079-Suhova_N:_Layout 1 18.05.2008 16:21 Page 514

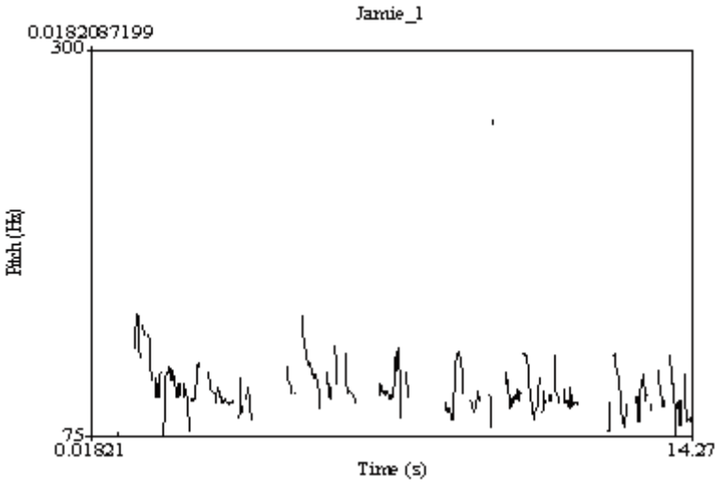


Рис. 1. Мелодический контур отрывка из эпизода Jamie_1.

В данном отрывке паузы хезитации выделяются на основании формальных характеристик, а также в силу того, что они зачастую соседствуют с синтаксической (межсинтагменной) паузой, употребляясь в пред-, пост- и интерпаузальной позиции (Бубнова 1994: 11). Так, в случаях (1, 2, 5a) – это *предпаузальное* употребление, т. е., другими словами, пауза колебания совпала с синтаксической (в нашем случае межсинтагменной) паузой; в примерах (4, 6) – *интерпаузальное*, т.е. пауза колебания встретилась в отрезке между двумя синтаксическими паузами; а в случае (3, 5b) – *постпаузальное*, т. е. на отрезке, следующим сразу за межсинтагменной паузой.

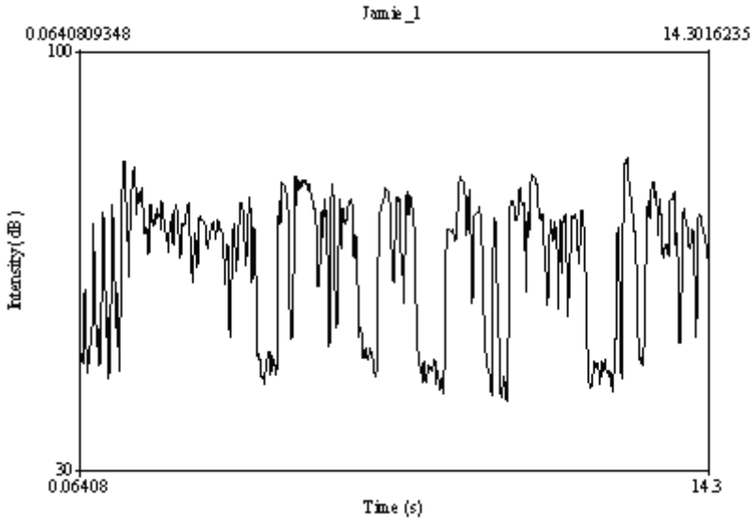


Рис. 2. Кривая интенсивности отрывка из эпизода Jamie_1.

Рассмотрим конкретные акустические характеристики в случаях пауз колебания в эпизоде *Jamie_1*.

(1) least | to:

Длительность отрезка паузы (между словами *least* и *to*) – 0,481 с.

Частота основного тона не меняется. Поэтому максимальное и минимальное, а значит, и среднее, значения частоты основного тона данного участка (участка паузы) определить не удалось.

Среднее значение интенсивности участка – 50 дБ (Рис. 3).

079-Suhova_N:_Layout 1 18.05.2008 16:21 Page 515

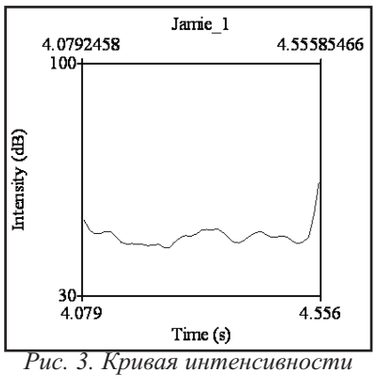


Рис. 3. Кривая интенсивности отрезка паузы (1).

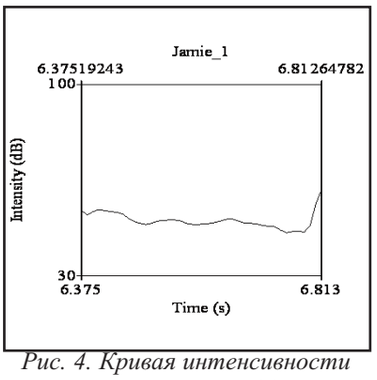


Рис. 4. Кривая интенсивности отрезка паузы (2).

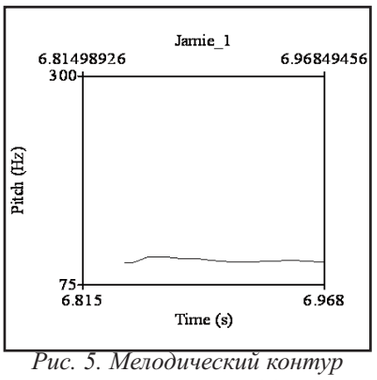


Рис. 5. Мелодический контур отрезка паузы (3).

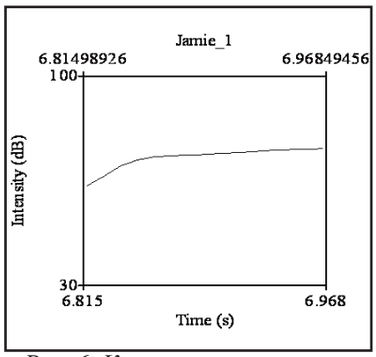


Рис. 6. Кривая интенсивности отрезка паузы (3).

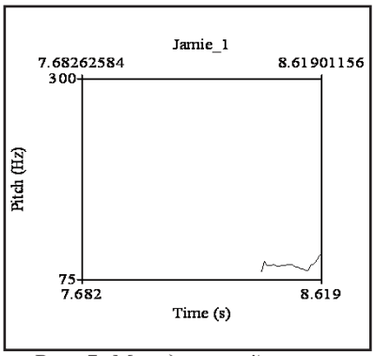


Рис. 7. Мелодический контур отрезка паузы (4).

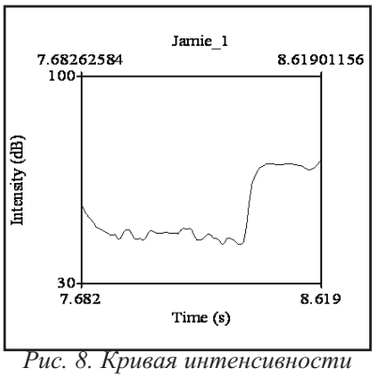


Рис. 8. Кривая интенсивности отрезка паузы (4).

(2) change⏐in:

Длительность отрезка паузы (между словами *change* и *in*) – 0,439 с.

Нет среднего значения частоты основного тона, средняя интенсивность – 51 дБ (Рис. 4).

(3) in (**3**)⏐not:

В данном случае, на самом деле, паузы между словами нет. Однако она существует, но не только на перцептивном уровне. На акустическом уровне происходит удлинение гласного [i], который звучит 0,153 с. Среднее значение частоты основного тона – 101 Гц (Рис. 5), средняя интенсивность –73 дБ (Рис. 6).

(4) ⏐***er***:

Итак, эта пауза состоит из перерыва фонации – паузы, длительность которой – 0,658 с, – и заполнителя-вокализации (*er*). Будем рассматривать весь отрезок. Он длится 0,936 с. Среднее значение частоты основного тона – 90 Гц (Рис. 7), средняя интенсивность – 64 дБ (Рис. 8).

(5) **(a)⏐*but* (b) ⏐*you know:***

Это сложный случай паузы колебания, общая продолжительность которой – 0,782 с, так как она состоит из двух частей: 1) пауза (между *school* и *but*), длительностью 0,207 с, и заполнитель *but*; 2) пауза (между *but* и *you know*), длительностью 0,278 с, и вводная конструкция *you know*. Однако весь этот отрезок можно считать паузой колебания между 2 синтагмами *in not just one school* и *my dream is to take over a borough*.

Целесообразным представляется посмотреть сначала на обе части по отдельности, а потом – вместе.

Пауза (1). Средняя частота основного тона – 188 Гц, средняя интенсивность – 64 дБ.

Пауза (2). Средняя частота основного тона –100 Гц, средняя интенсивность –70 дБ.

Общая пауза. Средняя частота основного тона –123 Гц (Рис. 9), средняя интенсивность –68 дБ (Рис. 10).

(6) over⏐a borough*:*

Эта пауза состоит из перерыва фонации – паузы, длительность которой – 0,629 с, – и удлинения гласного [ə]. Будем рассматривать весь отрезок. Он длится 0,818 с. Среднее значение частоты основного тона – 92 Гц (Рис. 11), средняя интенсивность – 58 дБ (Рис. 12).

079-Suhova_N:_Layout 1 18.05.2008 16:21 Page 516

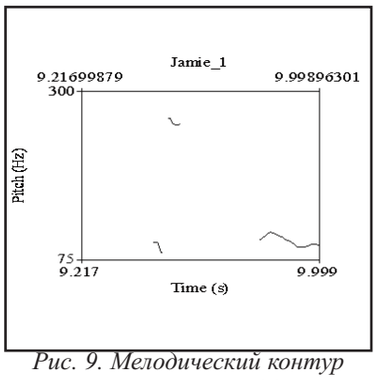


Рис. 9. Мелодический контур отрезка паузы (5).

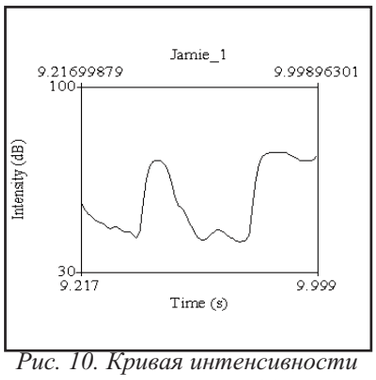


Рис. 10. Кривая интенсивности отрезка паузы (5).

Рис. 11. Мелодический контур отрезка паузы (6).

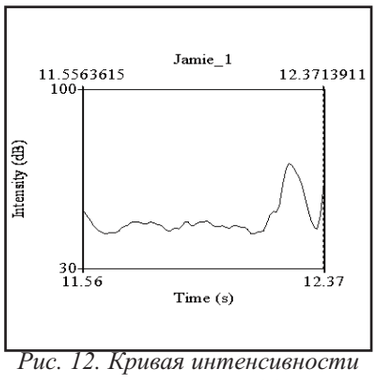


Рис. 12. Кривая интенсивности отрезка паузы (6).

Теперь попытаемся определить выделенные кинетические фразы, или, возможно, только отдельные кинетические жесты, которые сопутствуют участкам пауз колебания, будь то заполненные или незаполненные паузы, т. е. именно тем участкам, что описаны выше. В этой связи нас, на данном этапе, интересует выделенность или невыделенность того или иного жеста. Сводные данные по характеристикам пауз и жестов можно представить следующим образом (см. таб. 2).

Таким образом, видим, что в (1) на участке паузы колебания функционирует кинетическая фраза, состоящая из жестов *рук*, движений *головы*, значимых изменений *выражения лица*, *телодвижений* и *поз*. В других примерах не все кинетические формы задействованы в кинетических фразах. В основном, на участках пауз встречаются изменения положений *головы* и изменения *выражения лица*.

Отрезок речи (пример)	**Паузы[11]**			**Жесты**				
	Д (с)	**ЧОТ (Гц)**	**Инт (дБ)**	**Р**	**Г**	**ВЛ**	**ТД**	**П**
(1)	0,481	0	48	+ [12]	+	+	+	+
(2)	0,439	0	51	-	+	+	-	+
(3)	0,153	101	50	-	-	-	-	-
(4)	0,936	90	64	-	+	+	-	-
(5a)	0,207	188	64	-	+	+	-	-
(5b)	0,278	100	70	-	+	+	-	-
(6)	0,818	92	58	-	+	+	-	-

Таблица 2. Функционирование кинетических фраз на участках пауз колебания.

Несмотря на то, что это пилотный эксперимент, и проведен он на весьма небольшом речевом отрезке, основная его цель достигнута: можно предположить релевантные направления изучения способов взаимодействия пауз колебания и сопровождающих их кинетических фраз, а именно:

1) состав кинетических фраз;
2) степень выделенности той или иной кинетической формы в составе кинетической фразы;
3) вид паузы (пред-, интер- и постпаузальная);
4) тип пауз (заполненная / незаполненная);
5) виды заполнителей (фонетические и лексико-синтаксические заполнители);
6) показатели длительности пауз (длинные и короткие; усредненные значения);

[11] Отмечены такие характеристики как длительность (Д), частота основного тона (ЧОТ) и интенсивность (Инт).
[12] Укрупненные и смещенные влево значки указывают на то, что эти формы значительно выделены, по сравнению с другими.

079-Suhova_N:_Layout 1 18.05.2008 16:21 Page 517

7) показатели частоты основного тона и интенсивности (для заполненных пауз);

Представляется, что дальнейшее исследование взаимодействия пауз колебания и невербальных средств должно учитывать эти параметры и выявлять особенности разных способов их совместного функционирования с учетом этих характеристик.

Список литературы

1. Александрова О.А. Речекоммуникативный статус паузы колебания. Дис. ... канд. филол. наук. Великий Новгород, 2004. 208 с.

2. Александрова О.А., Иваницкий В.В. Пауза колебания – комплексный феномен современный коммуникации // Вестник новгородского университета, 2003. № 25. С. 95-101.

3. Бубнова Г.И. Просодия речевых «огрехов» // Экспериментальные исследования речи. Сб. научных трудов. М.: РАН, Институт языкознания, 1999. С. 5-48.

4. Дараган Ю.В. Функции слов – «паразитов» в русской спонтанной речи // Материалы международной конференции Диалог, 2000 / http://www.dialog-21.ru/materials/archive.asp?id=6260&y=2000&vol=6077 (28.01.08).

5. Дараган Ю.В. Паразитизм или симбиоз: механизм преодоления коммуникативных сбоев и обслуживающие его вербальные средства // Материалы международной конференции Диалог, 2003 / http://www.dialog-21.ru/Archive/2003/Daragan.htm (28.01.08).

6. Крейдлин Г.Е. Кинесика // Григорьева С.А., Григорьев Н.В., Крейдлин Г.Е. Словарь языка русских жестов. М.-Вена, 2001. С. 166-254.

7. Рыбка И.Н. Психолингвистическое исследование особенностей устного научного монологического текста. Автореф. дисс… канд. филол. наук. Уфа, 2007. 22 с.

8. Сухова Н.В. Взаимодействие просодического ядра и кинетической фразы в разных коммуникативно-прагматических типах монологических высказываний // Московский лингвистический журнал, 2006. № 9/1. С. 51-67.

9. Cecot M. Pauses in simultaneous interpretation: a contrastive analysis of professional interpreters' performances // The Interpreters' Newsletter, 11, 2001. P. 63-85.

10. Cerrato L., D'Imperio M. Duration and tonal characteristics of short expressions in Italian // Proceedings of the 15th International Congress of Phonetic Studies, 3-9 August, 2003. P. 1213-1216.

11. Proceedings of the 15th International Congress of Phonetic Studies, 3-9 August, 2003. P. 1209-1212.

12. Zellner B. Pauses and the temporal structure of speech // E. Keller (Ed.) Fundamentals of speech synthesis and speech recognition. Chichester: John Wiley, 1994. P. 41-62.

080-Tihomirov_I:_Layout 1 18.05.2008 16:21 Page 518

ИНТЕГРАЦИЯ ЛИНГВИСТИЧЕСКИХ И СТАТИСТИЧЕСКИХ МЕТОДОВ ПОИСКА В ПОИСКОВОЙ МАШИНЕ «EXACTUS»

INTEGRATION OF LINGUISTIC AND STATISTIC SEARCH METHODS IN SEARCH ENGINE "EXACTUS"

Тихомиров И.А. *(matandra@isa.ru),* ***Смирнов И.В.*** *(ivs@isa.ru)*
Институт системного анализа РАН

Доклад посвящен проблемам использования лингвистических методов поиска в современных поисковых машинах. Приведены особенности поискового алгоритма Exactus, представлена его экспериментальная оценка. Сделаны выводы о преимуществах объединения лингвистических и статистических алгоритмов поиска.

Введение

Разработки в области технологий поиска информации сосредоточены на методах, основанных на статистических характеристиках документов (TF*IDF-веса термов, ссылочное ранжирование и т.д.). Эти методы хорошо проработаны и успешно применяются во всех популярных поисковых машинах. С уверенностью можно утверждать, что дальнейшее развитие существующих статистических методов уже не может значительно улучшить качество поиска, в то время как остаётся ряд поисковых задач, с которыми указанные методы по своей природе не могут справиться. Среди этих задач поиск ответов на вопросы, поиск документов, близких по смыслу запросу, а также широкий класс задач, которые требуют рассмотрения текста как средства коммуникации, передачи смысла, а не как набора цепочек символов.

Между тем, коллективами лингвистов разработаны новые методы поиска и анализа текстов, которые в перспективе могут принести существенный выигрыш в точности и полноте поиска. [1, 2]. Основным препятствием к непосредственному использованию этих методов в поисковых машинах является сложность программной реализации и отсутствие экспериментальной проверки в условиях больших объемов данных. Неизвестно, насколько хорошо работает метод, пока он не проверен на больших коллекциях текстовых документов. Немаловажным является и тот факт, что коллективы лингвистов, как правило, не имеют хорошей аппаратной базы и опыта реализации задач в области программирования. Доведение лингвистического алгоритма «до ума» и его проверка в рамках серьезного соревнования (например, TREC или РОМИП) весьма трудоемкая, недешевая и, как следствие, неподъемная для лингвистов задача.

Серьезной проблемой является отсутствие у лингвистов знаний в области математики, что приводит к непониманию статистических формул и методов, используемых подавляющим большинством поисковых машин. В результате лингвистические алгоритмы никак не учитывают хорошо зарекомендовавшую себя статистическую составляющую алгоритмов поиска.

В последние несколько лет целью разработчиков поисковой машины Exactus является эффективное взаимодействие лингвистов, математиков, программистов на пути решения задачи объединения статистических и лингвистических методов поиска [1, 2]. В результате такого взаимодействия разработан экспериментальный прототип поисковой машины Exactus, сочетающий как статистические, так и лингвистические (синтаксис и семантику) подходы к поиску.

Особенности поискового алгоритма Exactus

Поисковый алгоритм Exactus объединяет статистические и лингвистические методы поиска. Из статистических характеристик текста в Exactus учитываются TF*IDF веса термов и значимость фрагментов текстов (на основе HTML-разметки документов). На этапе индексации документы преобразуются к формату «плоский текст», а затем в вектор слов документа с учетом морфологии русского языка. Оценка характерности слов документу рассчитывается на основе TF*IDF весов слов (термов). Далее, строятся прямой и обратный индексы слов, аналогично большинству поисковых машин Интернет [4].

Особенность Exactus заключается в том, что в индексах слова сортируются не только на основании их статистических характеристик, но и с учетом смысловых значений слов, которые определяются на основании

080-Tihomirov_I:_Layout 1 18.05.2008 16:21 Page 519

положений теории коммуникативной грамматики русского языка [3]. То есть слова рассматриваются не как лексемы, а как синтаксемы. Синтаксемой называется минимальная синтактико-семантическая единица языка, несущая свой обобщенный категориальный смысл в конструкциях разной степени сложности и характеризующаяся взаимодействием морфологических, семантических и функциональных признаков [3]. В основу подхода положено утверждение: смысл предложения определяется совокупностью входящих в него синтаксем и множеством связей на них. Такой подход позволяет разделять те же самые слова с точки зрения лексики в различных семантических значениях в индексе (субъект, объект, результатив и т.д.). Это, в свою очередь, позволяет более тонко сопоставлять поисковый запрос и документы в индексе, находя только те документы, в которые входят слова в том же семантическом значении, что и в запросе. Тем самым, в результатах поиска документы близкие запросу по смыслу выдаются раньше остальных, что принципиально невозможно достичь при использовании только статистических методов [2]. Рассмотрим пример:

Пусть имеется запрос: «К чему приводит гипертония?»,
и документы, содержащие следующие фрагменты текста:
Документ1: «Гипертония приводит к нарушению кровоснабжения тканей»,
Документ2: «Хроническое недосыпание приводит к гипертонии».

В запросе слово «что» в дательном падеже с предлогом «к» имеет семантическое значение «результатив» - результат воздействия чего-либо, а слово гипертония имеет семантическое значение «каузатив» - источник воздействия. В первом документе слово «нарушение», являющееся управляющим в словосочетании «нарушение кровоснабжения тканей», имеет семантическое значение «результатив» и будет являться ответом на поставленный вопрос. Во втором же документе, «результативом» является слово «гипертония», а «каузативом» слово «недосыпание», т.е. второй документ совсем не близок по смыслу запросу, даже напротив, хотя по словам достигнуто полное соответствие. Отсутствие лингвистических алгоритмов приводит к тому, что все популярные поисковые машины выдают результаты, схожие именно со вторым документом, в то время как по алгоритму Exactus первый документ более релевантен запросу.

Алгоритм поиска Exactus в индексе представляет собой слияние и пересортировку линейных упорядоченных списков, что является аналогичным концепции большинства поисковых машин [4]. Принципиальное отличие заключается в учете смысловых значений слов в индексе.

Экспериментальная оценка поискового алгоритма Exactus

Экспериментальная оценка алгоритма проводилась в рамках российского семинара по оценке методов информационного поиска в 2007 году [5]. Участникам семинара раздавались коллекции документов и запросов (несколько миллионов документов и несколько тысяч запросов). Коллекции документов индексировались, после чего по ним в автоматическом режиме прогонялись запросы. Результаты поиска помещались в файл, который затем обрабатывали независимые эксперты-оценщики, определяя степень релевантности запросов и документов. Методология оценки основывалась на следующих принципах:

1. Эксперт оценивает соответствие документов исходному запросу на основе расширенного описания информационной потребности (к каждому запросу прилагается краткое описание того, что должно быть по нему найдено).
2. Используется метод оценки типа «общего котла» (pooling) с глубиной пула 50 [5].
3. Используются следующие шкалы оценки релевантности:
   - точно релевантно;
   - возможно релевантно;
   - вероятно релевантно;
   - не релевантно;
   - невозможно оценить.
4. Результат считается релевантным, если он получил оценку по одному из первых двух пунктов шкалы.
5. Для выставления оценки результата используются два способа:
   - Строгая оценка AND – документ получает оценку релевантен или нерелевантен, если все оценщики выставили соответствующую оценку.
   - Нестрогая оценка OR - результат получает оценку релевантен, если хотя бы один оценщик выставил соответствующую оценку.

Для оценки используются метрики точности и полноты, а также 11-точечный график TREC, который отображает совмещенные показатели точности и полноты при разных показателях точности [5].

080-Tihomirov_I:_Layout 1 18.05.2008 16:21 Page 520

Результаты экспериментов Exactus приведены для поиска по коллекции «Белорусский WEB». Эта коллекция построена компанией Яндекс как выборка из страниц домена .by, присутствовавших в индексе поисковой машины Яндекс по состоянию на май 2007 года, объем коллекции 8 Гигабайт.

Ниже приводятся 11-точечные графики TREC для оценок AND и OR для системы Exactus и других участников семинара при поиске по коллекции «Белорусский WEB».

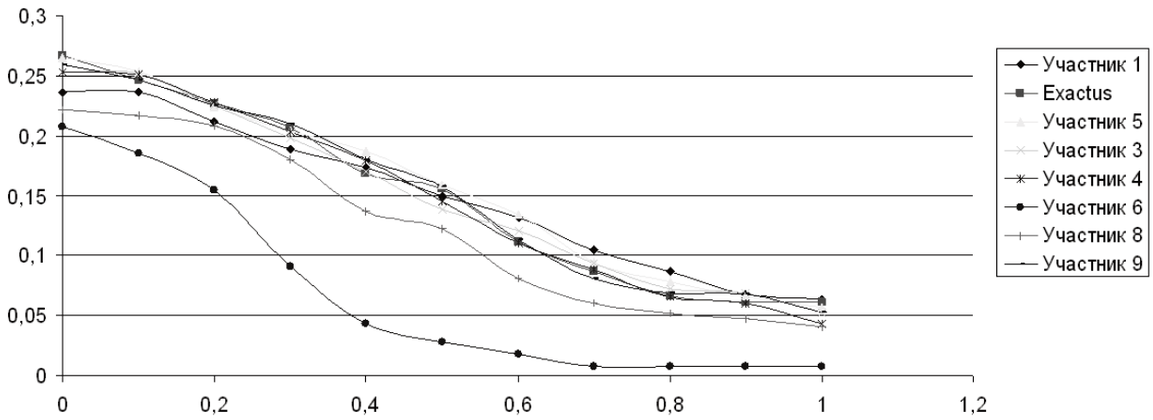


Рис. 1. График TREC: AND-оценка.

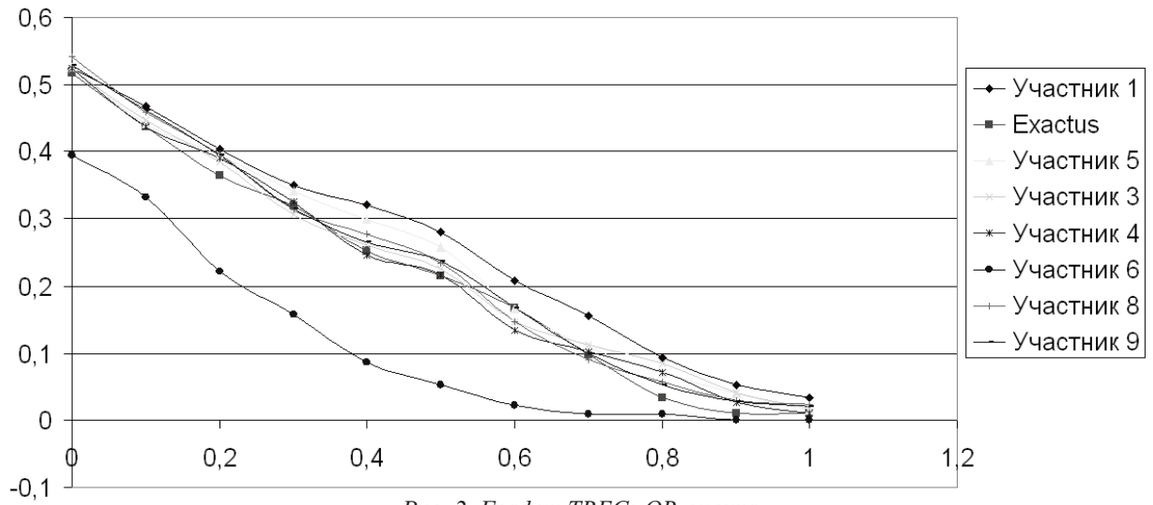


Рис. 2. График TREC: OR-оценка.

Как видно из графиков, алгоритм Exactus на коллекции «Белорусский WEB» показал результаты, которые в среднем не хуже результатов остальных участников. Следует отметить, что остальными участниками являются экспериментальные поисковые алгоритмы известных отечественных поисковых машин. Кроме того, в РОМИП в 2007 году использовалась свободно распространяемая поисковая машина Lucene (http://www.lucene.apache.org). Наилучшие результаты Exactus показал на строгой оценке AND, при высокой точности показав довольно высокие показатели полноты. Строгая оценка AND отражает мнения всех экспертов, т.е. более близка к независимой, объективной оценке.

Некоторое отставание от лидеров в оценке OR можно объяснить тем, что при поиске выполнялись алгоритмы семантического поиска, дающие приоритет смысловой составляющей текста, а не лексической, поэтому многие эксперты могли посчитать результаты не соответствующими словам запроса, т.е. не достаточно релевантными.

Заключение

Полученные в ходе экспериментов результаты показывают перспективность интегрирования лингвистических и статистических алгоритмов поиска текстов для задач повышения точности и полноты поиска.

080-Tihomirov_I:_Layout 1 18.05.2008 16:21 Page 521

Следует отметить тот факт, что поисковая машина Exactus показывает результаты, как минимум, на том же уровне, что и современные алгоритмы статистического поиска.

Важным результатом является и то, что применение серьезного арсенала лингвистических средств (включая синтаксис и семантику) для больших объемах данных позволило отработать Exactus на скорости, сопоставимой со скоростью работы современных поисковых машин Интернет. Это позволяет утверждать, что по-настоящему серьезные лингвистические анализаторы вполне применимы в реальных условиях. Тем более, что использование современных распределенных вычислительных систем и параллельных вычислений позволяет эффективно и в разумные сроки обработать большие коллекции текстовых документов.

Однако, всю силу лингвистических подходов к поиску возможно увидеть только на осмысленных запросах, являющихся полноценными высказываниями на естественном языке. В настоящее время сложилась культура поиска по ключевым словам, когда запрос задаётся перечислением значимых для пользователя слов, зачастую в словарной форме. Это, конечно, обесценивает использование лингвистики, однако с возрастанием объёмов информации в сети Интернет пользователю *придётся* выражать свою потребность в развёрнутом и более осмысленном виде, чтобы получать более релевантные его потребности результаты. Преимущества лингвистических подходов к поиску ответов на вопросы очевидны и не подлежат сомнению.

Дальнейшими направлениями исследований являются включение в алгоритм индексации ссылочного ранжирования и учет заранее составленного каталога ресурсов.

Список литературы:

1. Osipov G.S., Smirnov I.V., Tikhomirov I.A., Vybornova O.V, Zavjalova O.S. Linguistic Knowledge for Search Relevance Improvement.// Papers of Joint conference on knowledge-based software engineering JCKBSE‘06, IOS Press, 2006. - P. 294-302.

2. Осипов Г.С., Тихомиров И.А., Смирнов И.В. Exactus – система интеллектуального метапоиска в сети Интернет. // Труды десятой национальной конференции по искусственному интеллекту с международным участием КИИ-2006. М: Физматлит, 2006. т. 3. - С. 859-866.

3. Золотова Г.А., Онипенко Н.К., Сидорова М.Ю. Коммуникативная грамматика русского языка. Институт русского языка РАН им. В. В. Виноградова, М. 2004 – 544 с.

4. Sergey Brin, Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine. // http://infolab.stanford.edu/~backrub/google.html

5. Russian Information Retrieval Evaluation Seminar // http://romip.ru

081-Toldova:_Layout 1 18.05.2008 16:21 Page 522

СЕМАНТИЧЕСКИЕ ФИЛЬТРЫ ДЛЯ РАЗРЕШЕНИЯ МНОГОЗНАЧНОСТИ В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА: ГЛАГОЛЫ*

SEMANTIC FILTERS FOR THE WORD SENSE DISAMBIGUATION IN RNC: VERBS

Толдова С.Ю. *(toldova@yandex.ru)*
Московский государственный университет им. М.В. Ломоносова
Кустова Г.И. *(galina03@mtu-net.ru)*
Московский государственный педагогический университет
Ляшевская О.Н*. (olesar@mail.ru)*
Всероссийский институт научной и технической информации РАН

В статье обсуждаются результаты эксперимента по разработке системы семантических фильтров глаголов, используемых для разрешения неоднозначности лексико-семантической разметки в Национальном корпусе русского языка. Основные задачи эксперимента: проверить, в какой степени можно использовать специализированные лексикографические источники для создания таких фильтров (в качестве основного источника использовался словарь глагольного управления [Апресян-Палл 1982]); какие ограничения на актанты (семантические, лексические, грамматические) наиболее значимы для фильтров.

1. Введение

Настоящая статья продолжает серию публикаций в «Диалоге», освещающих работу над созданием лексико-семантической разметки Национального корпуса русского языка (http://www.ruscorpora.ru), см. [18], [19], [21], [22]. Все тексты Основного корпуса содержат три вида лингвистической разметки: метатекстовую (автор, жанр текста и т.д.), грамматическую (лемма и грамматические признаки) и лексико-семантическую (разметка по лексико-семантическим группам и словообразовательным типам). Сейчас на первый план выходит задача повышения точности разметки и снижения уровня «шума» в результатах поиска. Ее решение связано с учетом разных значений многозначных и омонимичных слов и с правильным распознаванием этих значений в тексте. В статье обсуждаются результаты экспериментов по разрешению неоднозначности семантической разметки глаголов.

В словаре Корпуса каждое значение слова снабжено семантическим ярлыком, показывающим его принадлежность к тому или иному таксономическому (семантическому) классу, например:

валяться

1) ‘движение: движение субъекта’ (*поросята валяются в грязи*);

2) ‘местонахождение’ (*бумаги валяются на полу*).

Таким образом, в словаре у многозначного слова обычно имеется несколько семантических помет, причем эти пометы распределены по разным значениям. Однако когда программа автоматически расставляет пометы в тексте, то она каждому вхождению слова приписывает все пометы, которые есть у слова в словаре (поскольку программа не может определить, в каком значении употреблено слово в каждом конкретном случае). В результате многозначное слово в тексте имеет все множество возможных помет. Это часто мешает более точному поиску в корпусе, создает «шум», а также иногда является источником ошибок морфологической разметки и лемматизации. В этой связи возникает задача найти способ снизить количество семантических помет для конкретного контекста без использования ручной разметки.

Для разрешения неоднозначности семантической разметки в Корпусе была разработана технология специальных фильтров. Семантический фильтр основан на принципе контекстной однозначности, т.е. на том, что в каждом конкретном контексте слово имеет одно значение (за исключением случаев языковой игры). Семантический фильтр – это правило, задающее некоторый минимальный контекст, в котором реализуется определенное значение слова. Если программа, содержащая фильтры, обнаруживает в предложении такой

* Работа выполнена при поддержке РГНФ, проект № 08-04-00181а.

Примеры взяты из Национального корпуса русского языка.

081-Toldova:_Layout 1 18.05.2008 16:21 Page 523

контекст, то она оставляет соответствующую ему семантическую помету, а остальные пометы удаляет. Таким образом, многозначность снимается с точностью до семантического класса (т.е. с точностью до семантической пометы). В фильтрах для разрешения неоднозначности прилагательных (см. [22]) используются только семантические характеристики определяемого существительного (поскольку грамматические характеристики существительного – род, падеж и число – не влияют на семантический класс (помету) прилагательного). В глагольных фильтрах тоже используются признаки связанных с глаголом существительных, но глагольные фильтры, в отличие от фильтров прилагательных, устроены более сложно. В них могут использоваться два параметра – семантические классы существительных, связанных с глаголом, и модель управления (МУ) глагола (т.е. морфолого-синтаксические характеристики существительных; учитываются также другие виды зависимых – придаточные, наречия). Мы исходили из того, что:

- глагол является синтаксическим и семантическим ядром предложения, а его базовые свойства определяются его МУ;
- значение глагола непосредственно связано с его МУ: с морфологическими (синтаксическими) и семантическими характеристиками его актантов (см. [6]).

В простейшем случае достаточно значения какого-то одного параметра – (1) модели управления глагола или (2) семантического класса существительного (далее МУ понимается в узком смысле – как «падежная рамка» глагола; возможно и широкое понимание МУ, когда в нее включаются не только грамматические, но и семантические характеристики актантов).

(1) Для идентификации значения может быть достаточно модели управления, если она является уникальной для данного значения. Например, у глагола *достать* в Корпусе (на уровне помет) различается три значения: ‘движение’ (*достать книгу с полки*), ‘обладание’ (*достать лекарство*, *достать билет в театр*) и ‘контакт’ (*достать рукой до потолка*). Если у первых двух значений модель управления может совпадать при неполной реализации (ср. *достать книгу* и *достать лекарство*), то последнее значение связано с уникальной моделью управления, и даже при ее неполной реализации (сущ.: Им. + *достать* + ***до*** сущ.: Род.) отличимо от первых двух. (2) Иногда для различения двух значений решающую роль играет семантическая характеристика актанта. Рассмотрим пример простейшего фильтра, различающего два значения глагола только за счет семантического класса управляемого существительного при совпадении модели управления. Многие глаголы физического воздействия имеют производное значение, относящееся к классу ‘речь’ (*пилить бревно* vs. *пилить мужа*, *резать хлеб* vs. *резать правду*, *молоть муку* vs. *молоть чушь*). Любое вхождение такого глагола в текстах Корпуса имеет две пометы – «физическое воздействие» (‘impact’) и «речь» (‘speech’). Фильтр содержит контекст, в котором реализуется одно из двух значений (контекст включает существительное с нужными грамматическими и семантическими характеристиками). Получая на вход такой контекст, программа оставляет у глагола нужную помету и автоматически удаляет ненужную:

(а) пилить (impact, speech) + сущ.: Вин.: конкр.: физич. предмет (*пилить бревно*) → пилить (impact);

(б) пилить (impact, speech) + сущ.: Вин.: конкр.: лицо (*пилить мужа*) → пилить (speech);

(а) молоть (impact, speech) + сущ.: Вин.: конкр.: вещество (*молоть муку*) → молоть (impact);

(б) молоть (impact, speech) + сущ.: Вин.: абстр.: речь (*молоть чушь*) → молоть (speech).

Конечно, в большинстве случаев ситуация намного сложнее.

Первая сложность связана с недостаточной различительной «мощностью» моделей управления. Часто у разных значений совпадают простейшие («минимальные») модели управления, включающие подлежащее и дополнение, ср.: *Он бросил снежок* vs. *Он бросил школу* vs. *Он бросил упрек.* Могут совпадать МУ с предложными группами: (а) *Он вернулся к столу* – (б) *Он вернулся к жене* – (в) *Сознание вернулось к нему* – (г) *Докладчик вернулся к первому вопросу*. Есть случаи, и их немало, когда совпадают не только минимальные, но и «расширенные» МУ: *Х бросил сумку на диван / в шкаф / за ширму* vs. *Х бросил взгляд на дверь / в окно / за ширму*; *Следователь вызвал свидетеля на допрос* vs. *Следователь вызвал свидетеля на откровенность*. В таких случаях нельзя обойтись только указанием МУ, необходимо включать в фильтр и семантическую информацию об актантах.

Другая сложность состоит в том, что количество именных групп в предложении, как правило, не совпадает с количеством именных групп, указанных в словарном источнике. В предложении могут содержаться именные группы, которые входят в состав других именных групп и не являются непосредственно актантами глагола: *Он нашел [для меня] [квартиру]* vs. *Он нашел [нож [для чистки картофеля]]*. С другой стороны, в реальном корпусе достаточно высок процент неполных предложений (около 10%), состоящих, например, только из одного глагола (ср. *Нашел.*). Мешают однозначно выделять актанты в реальном предложении и такие специальные конструкции, как комитативные и дистрибутивные группы, ср., например: *Он дал Пете по голове* vs. *Он дал каждому по прянику.*

Если говорить о семантических характеристиках, то здесь также возникает немало проблем. Во-первых,

081-Toldova:_Layout 1 18.05.2008 16:21 Page 524

существуют классы неодушевленных существительных, для которых характерны стандартные метонимические переносы, меняющие семантическую характеристику, например: организация → множество работающих в ней людей, ср. *Партия создана в 2001 г.* vs. *Партия решила…* Во-вторых, сложность в том, что иногда важно не противопоставление актантов по абстрактности/конкретности, а их объединение по некоторому семантическому компоненту ср. *Горит свет* (абстр. сущ.) и *Горит лампа* (конкр. сущ.). Кроме того, множества абстрактных и конкретных существительных неоднородны, поэтому иногда для различения значений необходимо выделять частные подклассы внутри класса абстрактных или конкретных существительных, ср., например: *Свет горит* vs. *План горит*.

2. Исходные данные

Каковы источники двух типов информации (МУ и семантические ограничения на актанты), используемой в фильтрах?

МУ можно извлекать как из текстов (из корпусов), так и из специальных и «обычных» словарей.

Задача выделения моделей управления (МУ) во многих системах автоматической обработки языка является актуальной как для синтаксического анализа, так и для разрешения семантической неоднозначности. Данная задача решается либо чисто статистическими способами (см. например, [1]), что приводит к потере точности, либо является трудоемким ручным процессом. Одним из способов преодоления указанных трудностей является создание специальных лексикографических ресурсов общего доступа таких, как WordNet, FrameNet и др. (см. [2], [3], [4], [5], [9]). Активно разрабатывается такой ресурс и для русского языка – RusNet – в группе под руководством И.В.Азаровой [15], [20]. Создание такого ресурса также достаточно трудоемкий процесс.

Что касается статистических методов, то для разрешения многозначности используются как контролируемые методы обучения, так и неконтролируемые ([1], [4], [8] и др., см. также обзор в [10]). Большинство таких систем базируются на баесовской модели или на модели канала с шумом. В работе [4] сообщается о достижении 90% точности для шести существительных с достаточно четко различимыми смыслами. Данный метод активно разрабатывался на материале английского языка. Он требует большого корпуса, размеченного вручную. Потенциальными признаками контекста являются все лексемы из достаточно большого окна. При неконтролируемом обучении (см., например [10]) невозможна семантическая разметка с приписыванием тому или иному слову семантического тэга, задача сводится к кластеризации множества контекстов на группы и их различение (discrimination).

Для задач нашего проекта было важно учесть опыт использования специализированных лексикографических ресурсов. Такие методы предполагают либо первичную полуавтоматическую разметку тренировочного корпуса (ср. проект Senseval [11]), либо использование тезаурусов и словарных систем, таких как Wordnet, FrameNet, VerbNet. Технологии применения данных систем активно разрабатываются в проектах по семантическому аннотированию корпусов на многих языках, в том числе при разрешении многозначности глаголов с использованием моделей управления: [6], [7], [12], [13], [14].

Для русского языка в рамках проекта RusNet проводились пилотные эксперименты по применению лексикографических источников для извлечения моделей управления (см. [20]). Однако данный проект предполагает задачу лексикографического описания глагола, а не снятие омонимии в корпусе.

В своей работе мы опирались на опыт группы разработчиков RusNet, однако наш эксперимент был призван оценить, каким образом можно использовать готовые лексикографические источники и каким образом дополнять извлеченную из этих источников информацию с использованием обучающего корпуса.

В качестве основного источника МУ глаголов использовался словарь глагольного управления Апресян-Палл 1982 [17]. Из словаря извлекалась информация о различных возможных наборах актантов и сирконстант для разных значений глагола, о грамматических ограничениях на них.

Что касается второго параметра, то в качестве источника информации о семантических ограничениях использовалась таксономическая разметка существительных НКРЯ. Первоначально учитывалась только минимальная семантическая и лексико-грамматическая информация об актантах: одушевленность / неодушевленность и абстрактность / конкретность. Несмотря на перечисленные в разделе 1 сложности, минимальная грамматическая и семантическая информация способна существенно снизить степень многозначности. Если минимального набора признаков оказывалось все-таки недостаточно, привлекалась более детальная информация о таксономическом классе соответствующих существительных. Имеющаяся в Корпусе семантическая разметка для целей эксперимента была дополнена новыми пометами, а именно: (а) была расширена система таксономических классов; (б) учитывались метафорические переносы (помета «metaph»); (в) для служебных значений (лексических функций в смысле [16], ср., например, *найти* в *найти возможность*) была введена помета «LF».

081-Toldova:_Layout 1 18.05.2008 16:21 Page 525

Для уменьшения ошибок, связанных с отсутствием синтаксического анализа, мы использовали преобразования исходного контекста, моделирующие неполный синтаксический анализ.

Материалом эксперимента послужил корпус со снятой морфологической омонимией объемом 4,5 млн. словоупотреблений. Исследовались глаголы из высокочастотной части списка.

Эксперимент должен был ответить на следующие вопросы:

- ✓ в какой степени можно использовать информацию автоматически или полуавтоматически извлеченную из лексикографических источников;
- ✓ в какой степени данные о МУ глагола с использованием минимальной информации о семантическом классе актантов (одушевленность vs. неодушевленность, абстрактность vs. конкретность) позволяют понизить степень многозначности;
- ✓ каким образом извлекать информацию об актантах глагола в конкретном контексте, не используя полный синтаксический анализ;
- ✓ какова технология пополнения исходного списка МУ с использованием обучающего корпуса;
- ✓ каков должен быть формат глагольного фильтра для разрешения семантической неоднозначности;
- ✓ как взаимодействуют различные таксономические, грамматические и лексические ограничения.

3. Использование информации о грамматических и семантических ограничениях на актанты при создании семантических фильтров для разрешения глагольной многозначности

3.1. Вклад информации о составе и грамматических характеристиках элементов МУ в разрешение глагольной многозначности

Сначала был проведен эксперимент, имеющий целью установить, каков вклад собственно грамматической информации об актантах в разрешение многозначности глагола.

Для каждого исследованного глагола составлялся тестовый корпус предложений с данным глаголом (в них встречались и полные МУ, соответствующие словарному источнику [17], и не полностью реализованные МУ, и вхождения глагола без распространителей). В качестве примера ниже приводится диаграмма 1, которая показывает распределение моделей управления глагола *давать* в Корпусе:

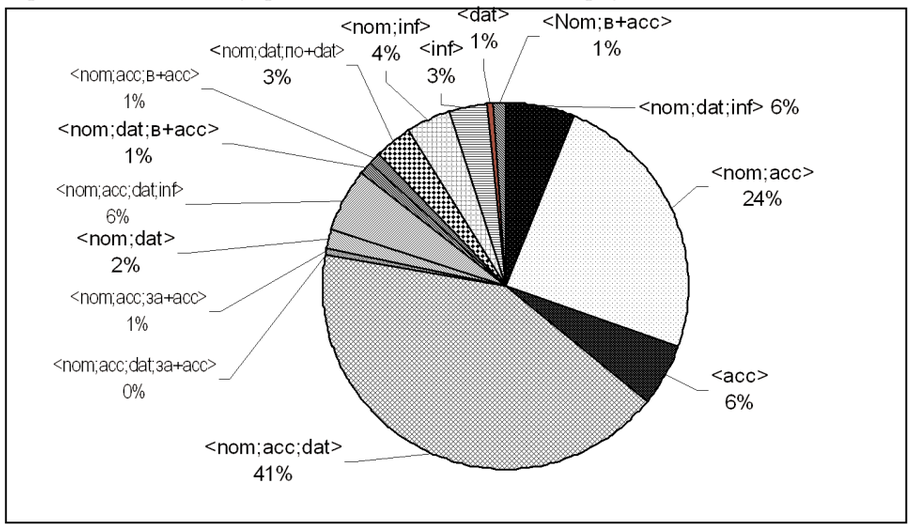


Диаграмма 1.

Как видно из диаграммы, МУ, включающие базовые актанты (<nom, acc, dat> и <nom, acc>), составляют бóльшую часть примеров корпуса.

Анализ тестового корпуса позволил выявить (1) случаи, препятствующие разрешению омонимии, и (2) случаи, способствующие ее разрешению.

(1) Факторы, повышающие неоднозначность.

(а) Реализована базовая МУ.

Показательным является тот факт, что базовая, «стандартная» МУ, характерная для данного глагола или класса глаголов, обычно обладает наибольшей степенью многозначности. Так, базовая МУ глагола *дать* / *давать* (и других глаголов этого класса) <именительный, винительный, дательный> представлена почти во всех

081-Toldova:_Layout 1 18.05.2008 16:21 Page 526

возможных для данного глагола значениях: прямое значение – класс ‘каузация обладания’ (*Мать дала ему пирожок*), лексические функции (*Мы дадим вам такую возможность*; *Войска дали отпор врагу*), класс ‘физическое воздействие’ (*Она дала ему пощечину* – это лексическая функция, однако она может быть «семантизирована»). При этом такая модель, как правило, имеет наибольшее покрытие (37% для глагола *дать*). Базовая модель <именительный, винительный> глагола *покинуть* также представлена в разных значениях: прямое значение – класс ‘движение’ (*Новобранцы покинули родное село*), лексическая функция (*Смелость покинула его* – ‘исчезновение’), фазовое значение (*Певица покинула сцену*).

(б) Модель управления реализована не полностью.

В корпусе встречается достаточно много случаев, когда некоторая МУ реализована не полностью или в доступном анализу отрезке текста представлен только глагол без актантов. Вот несколько примеров для глагола *думать*: *в министерстве действительно могут так думать*; *чтобы она не мешала мне думать*; *надо думать*; *я думаю*; *потому что думал*; *и думать не хочу*; *продолжал мучительно думать*; *а по-настоящему думать* и т.п.

(2) Факторы, понижающие неоднозначность.

(а) Модель управления, включающая «специфичные» актанты, существенно сужает число возможных значений вплоть до одного. Так, например, для глагола *болеть* предложная группа *за*+S&acc в МУ задает только одно значение: *Он болеет за «Динамо»*. Значение глагола *найти* в контексте прилагательного в творительном падеже относится к классу ‘ментальные действия’ или ‘восприятие’ (*Книгу я нашёл весьма грамотной*). Глагол *дать* при наличии предложных групп *в*+Вин. или *по*+Дат. реализует значение ‘физическое воздействие’ (*Здорово ему давеча Кирилл Анатольевич дал по башке*). Для глагола *толкать* актант *на*+S&acc в МУ задает только одно значение (*толкать на преступление*). Реализация валентности инструмента у «физического» значения глагола *пилить* (*пилить бревно* (Вин.) *пилой* (Твор.)) позволяет однозначно отличить его от речевого значения (*пилить мужа*). У речевого значения, в свою очередь, есть валентность мотивировки (*пилить за что*), которой тоже достаточно для его идентификации. Разное падежное оформление второго актанта при глаголах движения также позволяет существенным образом сузить класс значений. Так, глагол *идти* имеет по разметке НКРЯ 8 тэгов. Для значения ‘движение’ возможно более 20 МУ. Однако каждая из этих МУ либо связана только с данным значением, либо максимальная величина кластера не превышает 3-х значений.

Таким образом, МУ может быть надежным критерием для сужения значения: если в предложении помимо собственно синтаксических валентностей (соответствующих подлежащему и прямому дополнению) реализуются специфичные валентности, обусловленные особенностями семантики конкретного глагола, а также факультативные валентности или некоторые сирконстанты, учет этих распространителей нередко позволяет отличить одно значение от другого, не прибегая к семантическим признакам существительных.

(б) Отсутствие в реальном предложении каких-либо именных групп не обязательно ведет к повышению неоднозначности (к реализации всех или большинства возможных значений), для некоторых глаголов такой контекст, наоборот, снижает число возможных семантических тэгов вдвое. Например, для глагола *дать* МУ с отсутствием прямого дополнения в винительном падеже может сигнализировать о том, что реализовано значение ‘физическое воздействие’ (*А он ему как дал*); отсутствие актанта в дательном падеже характерно для некоторых лексических функций (*дать течь; дать свисток; дать эффект*).

(в) Неполная реализация МУ, ее редукция, вплоть до отсутствия синтаксически выраженных актантов в определенных конструкциях, также не всегда является негативным фактором, в некоторых случаях она может, наоборот, служить для разрешения неоднозначности, сужая число возможных значений. Так, употребление глагола *толкать* без падежных распространителей в неопределенно-личной конструкции (*Сзади толкают*) возможно только для первого (физического) значения.

3.2. Семантические ограничения

Следующим диагностическим признаком является семантический класс актанта. Однако данная характеристика играет роль диагностического признака далеко не всегда. Один и тот же семантический признак актанта для одних глаголов может быть решающим, а для других – ни о чем не говорить. Так, для глаголов движения прямое значение физического перемещения характерно как для одушевленных, так и для неодушевленных объектов, при этом и тот, и другой класс может участвовать в метафорических переносах и сочетаться с лексическими функциями (ср. *Поезд идет из Москвы ~ Человек идет из сада ~ Чай идет из Индии* и т.п.). Для глаголов же восприятия или ментальных глаголов наличие неодушевленного подлежащего очень маловероятно. А это значит, что, например, глагол *найти* не может реализовывать одно из своих переносных значений, относящихся к классу ментальных, в контексте, когда позицию подлежащего занимает неодушевленный актант (ср. *Метод нашел применение*…). Семантические ограничения в сочетании с синтаксической ролью образуют иерархию с точки зрения надежности отсечения лишних значений.

081-Toldova:_Layout 1 18.05.2008 16:21 Page 527

Абстрактность актанта чаще играет решающую роль в определении значения глагола, чем одушевленность. Так, для глагола *дать* абстрактность существительного в позиции прямого дополнения является решающим ограничением для выделения употреблений данного глагола как лексической функции. При этом абстрактность актанта, занимающего позицию подлежащего, более надежный признак, чем абстрактность локативного актанта.

Анализ данных показывает, что чем «специфичней» ограничения, тем точнее может быть разрешена многозначность. Иногда приходится прибегать к более частным семантическим признакам в рамках широких классов абстрактности / конкретности. Так, если в случаях *бросать гневные упреки* / *горькую правду* / *чудовищные обвинения* не ограничиваться пометой «абстр.» для сущ. в Вин., а использовать помету «речь», то данное значение глагола *бросать* тоже можно будет идентифицировать как «речь» (а не как бессодержательную «лексическую функцию»). В случаях *оторвать голову от подушки*, *не отрывать глаз от книги* общую характеристику существительного в Вин. «конкр.» имеет смысл дополнить более частной характеристикой «часть тела» (впрочем, эта помета может использоваться для идентификации данного значения только совместно с грамматической характеристикой другого актанта «*от* + сущ: Род.», т.к. семантическая характеристика «часть тела» может быть у актанта и в другом значении, ср.: *взрывом оторвало ногу*).

В некоторых случаях приходится даже использовать лексические фильтры, т.е. правила, в которых фигурируют конкретные лексемы. Например, для глагола *болеть* важно, что отдельно необходимо рассматривать словосочетание *болеть* с существительным *душа*: *болеть душой* – наличие в предложении слова *душа* однозначно указывает на непрямое, метафорическое значение глагола. В контексте данного существительного значение глагола следует отнести к классу ‘выражение эмоций’. Т.е. почти со 100% точностью можно во всех подобных примерах оставить ровно одно значение.

3.3. Некоторые результаты эксперимента

Эксперимент показал, что грамматические характеристики актантов и сирконстант позволяют существенным образом понизить многозначность глаголов. Особенно информативны оказываются более периферийные актанты. При этом можно разбить глаголы на классы в зависимости от того, в какой степени именно грамматическая информация позволяет уменьшать число возможных значений. Что касается лексико-грамматических и семантических характеристик, то самых общих признаков «одушевленность» / «неодушевленность» и «конкретность» / «абстрактность» иногда оказывается достаточно для существенного понижения степени многозначности.

Рассмотрим диаграмму 2, в которой отражены свойства грамматических и обобщенных семантических ограничений для некоторых глаголов.

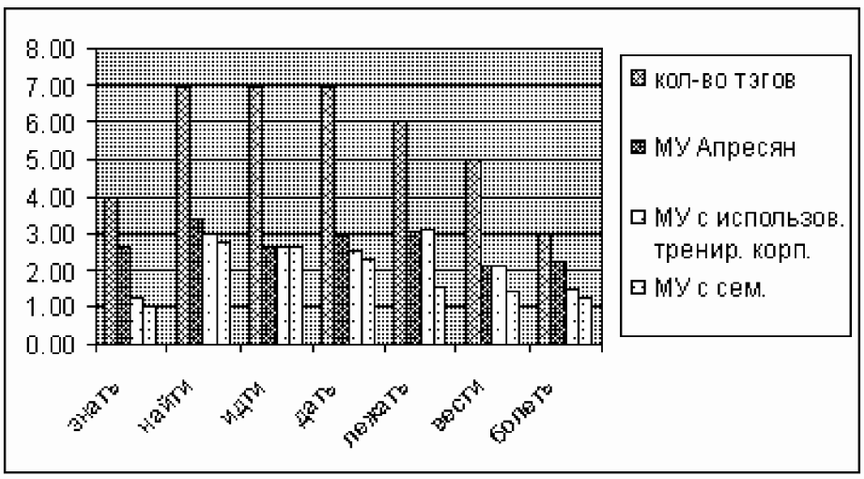


Диаграмма 2.

Для глаголов *найти*, *идти*, *дать, лежать* информация о грамматических свойствах актантов позволяет снизить число возможных значений более чем в два раза. При этом использование корпусных данных в ряде случаев существенно улучшает результаты применения грамматических фильтров (ср., например, данные для глаголов *знать, болеть*). Как видно из диаграммы, семантические ограничения также имеют разное значение для разных классов глаголов. Так, включение в число ограничений обобщенных семантических характеристик

081-Toldova:_Layout 1 18.05.2008 16:21 Page 528

актантов глагола *идти* совсем никак не влияет на снижение степени многозначности. Для глаголов же *лежать, вести, болеть* такие характеристики позволяют снизить многозначность почти до одного тэга на глагол, т.е. полностью разрешают многозначность в большинстве контекстов.

4. Выводы

Как синтаксические характеристики актантов, так и семантические ограничения на них могут иметь разную различительную силу.

С точки зрения различительной силы актанты образуют иерархию. Их можно разбить на два класса:

(а) базовые актанты, такие как S&nom, S&acc, а также актанты, соответствующие семантическому классу глагола в его прямом значении (например, актант, указывающий на место, для глаголов движения; датив для глаголов класса *давать*);

(б) уточняющие актанты.

МУ, содержащая базовые актанты, приводится первой в словарных источниках. Базовые актанты наиболее частотны при данном глаголе в корпусе (более 60%). Они, как правило, содержатся в нескольких МУ данного глагола и реализуются с несколькими значениями.

Второй класс актантов включает более специфические, необязательные актанты, например, предложные группы (*за*+S&acc, *по*+S&dat и др.) или инфинитив. Они обладают большей «различительной» силой. Наличие такого актанта в МУ существенным образом сужает множество соответствующих значений вплоть до одного, таким образом, он может служить диагностическим признаком для некоторого значения даже при отсутствии в контекстной МУ других актантов.

Описанное выше противопоставление «различительных» и «неразличительных» актантов вполне предсказуемо. Неожиданным результатом явился тот факт, что для многих глаголов ситуация, когда по некоторым причинам в предложении не хватает актантов, оказывается также более «благоприятной» для разрешения многозначности, чем полная стандартная модель. Так, например, в предложении *Он дал* глагол *дал* с достаточно высокой степенью вероятности не может иметь ни значения лексической функции, ни значения физического воздействия, ни значения ‘позволить’, а относится к исходному классу ‘каузировать иметь’. Множество значений глагола *найти* в предложении с опущенными актантами также уменьшается с пяти максимально возможных до двух (обладание (*найти кошелек*) и метафорический перенос по этому значению (*найти выход*), ср.: *Он долго искал кошелек* / *выход. И наконец нашел*). Подобную информацию можно извлечь только из размеченного обучающего корпуса, либо опираясь на интуицию эксперта, поскольку никакие лексикографические источники такой информации не дают по определению. Это не входит в их задачу.

Наибольшую сложность для снятия многозначности представляют случаи, когда для разных значений набор основных актантов совпадает. В такой ситуации признаки образуют некоторую иерархию с точки зрения их различительной силы. Наибольшей степенью различительности обладают лексические ограничения (случаи, когда глагол реализует данное значение только в устойчивом словосочетании), далее следуют периферийные (факультативные актанты), а также такие семантические характеристики, как абстрактность.

Список литературы

1. Brown, P.F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert Mercer. Word-sence disambiguation using statistical methods. // ACL. 1991. V.29. P. 264–270.

2. Dagan I., Itai A., Schwall U. Two languages are more informative than one // Proceedings of the ACL, 1991 (29). P. 130–137.

3. Fellbaum, Christian (ed.) WordNet: An Electronic Lexical Database. MIT Press. 1998.

Gale, William A., Church, Kenneth W. and Yarowski, David. A method for disambiguating word senses in a large corpus. // Computers and the Humanities. 1992. Vol. 26. P. 415–439.

4. Gildea, Daniel, Daniel Jurafsky. Automatic Labeling of Semantic Roles // Computational Linguistics. 2002. Vol. 28. No 3. P. 245–288.

5. Johnson,C., Fillmore,C., Petruck,M., Baker,C.,Ellsworth,M., Ruppenhofer, J., and Wood, E. FrameNet: Theory and Practice. 2002. [Electronic resource]. 2002. Mode of access: http://www.icsi.berkeley.edu/framenet.

6. Kingsbury, P., Palmer, M., and Marcus, M. Adding semantic annotation to the Penn TreeBank. // Proceedings of the Human Language Technology Conference HLT-2002. San Diego, California, 2002.

7. Lesk M. Automatic sense disambiguation using machinereadable dictionaries: How to tell a pine cone from a ice cream cone. // Proceedings of SIGDOC ’86. New York. Association for Computing Machinery. 1986. P. 24–26.

081-Toldova:_Layout 1 18.05.2008 16:21 Page 529

8. Lopatková, Markéta, Ondřej Bojar, Jiří Semecký, Václava Benešová, and Zdenék _Zabokrtský. Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation. // Václav Matoušek, Pavel Mautner, and Tomáš Pavelka, editors. Text, Speech and Dialogue: 8th International Conference, TSD 2005. – Karlovy Vary, Czech Republic, September 12-15, 2005. Proceedings, volume LNAI 3658. Springer Verlag. 2005. P. 99–106.

9. Manning C.D., Schütze H. Foundations of Statistical Natural Language Processing. Chapter 7. Cambridge, Massachusetts: The MIT Press. 1999. P.230–262.

10. Mihalcea R., Chklovsky T., Kilgarriff A. Framework and results for English SENSEVAL // Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, July 2004, Barcelona. Barselona, Spain, 2004. P. 25–28. ftp://ftp.itri.bton.ac.uk/reports/ITRI-04-09.pdf.

11. Ng H.T., Lee H.B. Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach // Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96). Santa Cruz, 1996.

12. Scott Songlin Piao, Rayson P., Archer D., McEnery T. Comparing and combining a semantic tagger and a statistical tool for MWE extraction // Computer Speech & Language. Vol. 19. No 4. 2005. P. 378–397.

13. Shi, L., and Mihalcea, R. Semantic parsing using FrameNet and WordNet. // Proceedings of the Human Language Technology Conference (HLT/NAACL 2004). Boston, 2004.

14. Азарова И.В., Синопальникова А.А., Яворская М.В. Принципы построения wordnet-тезауруса RussNet // Кобозева И.М., Нариньяни А.С., Селегей В.П. (ред.), Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог'2004. М.: 2004. С. 542–547.

15. Апресян Ю.Д. Лексическая семантика. – М.: «Наука», 1974. – 368 с.

16. Апресян Ю.Д., Палл Э. Русский глагол – венгерский глагол. Управление и сочетаемость. Будапешт, 1982.

17. Кобрицов Б.П., Ляшевская О.Н., Шеманаева О.Ю. Снятие лексико-семантической омонимии в новостных и газетно-журнальных текстах: поверхностные фильтры и статистическая оценка // Интернет-математика – 2005. М.: 2005. С. 38–57.

18. Кустова Г.И., Ляшевская О.Н., Падучева Е.В., Рахилина Е.В. Опыт семантического расширения морфологической разметки: таксономическая классификация лексики в национальном корпусе русского языка // НТИ, сер. 2. Информационные процессы и системы. № 6. 2005.

19. О.А. Митрофанова, В.В. Кадина, В.С. Савицкий. Экспериментальное исследование синтагматических свойств лексем на основе лексикографических описаний и корпусов текстов // Труды международной конференции MegaLing'2006–Горизонты прикладной лингвистики и лингвистических технологий. 20–27 сентября 2006 г., Украина, Крым, Партенит.

20. Рахилина Е.В., Ляшевская О.Н., Кобрицов Б.П., Кустова Г.И., Шеманаева О.Ю. Многозначность как прикладная проблема: Лексико-семантическая разметка в Национальном корпусе русского языка // Лауфер Н.И., Нариньяни А.С., Селегей В.П. (ред.). Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006». 2006. С. 445–450.

21. Шеманаева О.Ю, Кустова Г.И., Ляшевская О.Н., Рахилина Е.В. Семантические фильтры для разрешения многозначности в Национальном корпусе русского языка: прилагательные // Иомдин Л.Л., Лауфер Н.И., Нариньяни А.С., Селегей В.П. (ред.). Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007». 2007. С. 582–587.

082-Uryson_E:_Layout 1 18.05.2008 16:24 Page 530

СОЮЗЫ *А ТО* И *А НЕ ТО*: ПОЧЕМУ В НЕКОТОРЫХ КОНТЕКСТАХ ОНИ СИНОНИМИЧНЫ*

RUSSIAN CONJUNCTIONS *A TO* [LIT.: 'AND/BUT THAT'] AND *A NE TO* [LIT.: 'AND/BUT NOT THAT']: WHY ARE THEY SYNONYMS IN SOME CONTEXTS?

Урысон Е.В. *(x-uryson@mtu-net.ru)*
Институт русского языка им. В.В. Виноградова РАН

Союзы *а то* и *а не то* формально различаются на отрицание (не), однако в некоторых контекстах они синонимичны. Причина в том, что в состав этих союзов входит анафорический элемент *то*. В союзе *а то* данный элемент отсылает к одному фрагменту семантической структуры фразы, а в *а не то* – к другому. Обсуждаются возможности композиционального анализа союзов и частиц.

1. Введение

1.1. Объект исследования. Предмет нашего исследования – союзы *а то* и *а не то*, представленные в контекстах типа:

(1) *Купались, ходили за грибами, а то <а не то> брали лодку и уплывали на соседний остров.*

(2) *Отдай машинку, а то <а не то> маме пожалуюсь*.

В подобных контекстах союзы *а то* и *а не то* синонимичны и взаимозаменимы.

Заметим, что в этих же контекстах может выступать и союз *не то*. Ср.

(1а) *Купались, ходили за грибами, не то брали лодку и уплывали на соседний остров.*

(2а) *Отдай машинку, не то маме пожалуюсь*.

Союз *не то* мы не рассматриваем за недостатком места. В частности, мы оставляем в стороне вопрос, является ли *не то* просто усеченным вариантом союза *а не то* или же это отдельное слово. Однако основные выводы относительно *а то* и *а не то* распространяются и на союз *не то*.

Союзы *а то* и *а не то* многозначны. В частности, в примере (1) они представлены в одном значении, а в (2) – в другом. Иными словами, в этих примерах фигурируют их разные лексемы (подробнее см. ниже). У данных союзов есть и другие значения, причем в этих значениях *а то* и *а не то* не взаимозаменимы. Ср. следующие примеры:

(3) *Пойду, а то <*а не то> уже поздно*.

(4) *Он теперь старый, больной, полуслепой, а то <*а не то> был сильный, властный, богатый*.

Мы ограничимся анализом тех лексем *а то* и *а не то*, которые представлены в (1)-(2). За пределами нашего описания остаются, в частности, единицы *не то чтобы*, *не то что* т.п., включающие в свой состав цепочку *не то*. Ср. *Она не то чтобы обиделась, но очень расстроилась*; *Он не то что злой, а какой-то раздраженный*. Наконец, мы не рассматриваем сочетание *а не то*, в котором представлено указательное местоимение *то* в контексте противопоставления. Ср. *Делай то, что требуют, а не то, что хочется*.

1.2. Постановка вопроса. В традиционной грамматике союзы принято классифицировать, в частности, по их строению. С этой точки зрения союзы *а то* и *а не то* являются составными, т.е. неоднословными: они «представляют собой соединение двух или более элементов, каждый из которых одновременно существует в языке и как отдельное слово» [Грамматика-80: 716]. При этом данные союзы различаются всего на один элемент – на отрицание *не*.

Казалось бы, семантически союзы *а то* и *а не то* должны различаться так же, как и другие сочетания вида *Р – не Р*. Ср. *Спит – Не спит, в лесу – не в лесу* и т.п. В самом общем виде, частица *не* в подобных случаях выражает смысл 'неверно, что Р'. Ясно, однако, что союзы *а то* и *а не то* если и различаются в рассматриваемых значениях, то не компонентом 'неверно, что'.

Быть может, данные союзы противопоставлены по смыслу не как словосочетания, а как отдельные слова вида *Х – неХ*? Ср. *ловкий – неловкий, красивый – некрасивый*, *старый – нестарый* (ср. *еще нестарая женщина*)

* Работа выполнена при поддержке РГНФ (проект № 06-04-00289а), программы фундаментальных исследований ОИФН РАН «Русская культура в мировой истории» и гранта Президента РФ НШ-3205.2008.6

082-Uryson_E:_Layout 1 18.05.2008 16:24 Page 531

и т.п. В общем случае слова в подобных парах являются антонимами.

Парадоксальным образом союзы *а то* и *а не то* в рассматриваемых значениях не только не являются антонимами, но даже воспринимаются как абсолютные синонимы – они свободно взаимозаменимы без сколько-нибудь заметного изменения смысла высказывания. Этот факт требует объяснения. Требуется описать данные лексемы союзов *а то* и *а не то* так, чтобы из описания стало ясно, почему они синонимичны, а не антонимичны, как казалось бы, естественно было ожидать.

Нужно сказать, что союзы *а то* и *а не то* неоднократно привлекали к себе внимание лингвистов и их семантика уже в определенной степени описана [Санников 1989; Белошапкова 1970; Подлесская 2000; Собинникова 1969; Колосова 1980; Инькова-Манзотти 2000; Израэли 2000]. Однако нужные нам сведения невозможно извлечь из толкования лексем союзов *а то* и *а не то*. Для ответа на поставленный вопрос требуются иные подходы.

1.3. Метод исследования. Обратим еще раз внимание на то, что союзы *а то* и *а не то* состоят из хорошо выделимых элементов: *а*, *то* и *не*, которые сами могут функционировать как отдельные слова. В этом отношении союзы *а то* и *а не то* отнюдь не уникальны. В славянских языках, в частности в русском, довольно много союзов и частиц, состоящих из элементарных компонентов, которые тоже являются служебными словами. Ср. *или = и+ли*; *чтобы = что+бы*; *дабы = да+бы*; *все же = все+же*; *а также = а+так+же*; *даже = да+же* и т.п. Правда, в некоторых таких «макромолекулах» выделяются элементы, которые в современном русском языке уже не функционируют в качестве отдельных слов. Однако соответствующие слова обнаруживаются в истории языка, в русских диалектах или в других славянских языках. Ср. *ибо = и+бо*, *либо = ли+бо*. В современном русском литературном языке частицы *бо* нет, однако она фиксируется в памятниках, в некоторых диалектах, в украинском языке.

На такие «макромолеуклы» неоднократно обращала внимание Т.М.Николаева [Николаева 1982; 2003]. Гипотеза Т.М.Николаевой состоит в следующем. В праславянском языке и в разных славянских языках на достаточно ранней стадии их развития отдельные «партикулы» (т.е. частицы, союзы и вообще элементы с подобной семантикой) соединялись в новые, составные партикулы. Так формировался современный фонд славянских партикул. На какой-то стадии развития славянских языков механизм склеивания партикул уже не работал.

Нас интересует семантическая сторона этого механизма. Предлагаемая нами гипотеза состоит в следующем.

Соединение двух или более партикул в «макромолекулу» сопровождалось как минимум двумя разными семантическими процессами. Это, во-первых, десемантизация, т.е. семантическое выветривание, опустошение элементов, соединяющихся в новую единицу. Во-вторых, это сплавление, спекание элементов внутри «макромолекулы». В результате этих процессов получившаяся «макромолекула» представляет собой действительно новое слово, а не просто сочетание двух или более слов. Ср. *даже*; в составе этой частицы выделяются две «партикулы»: *да* и *же*. Элемент *да* по форме совпадает с частицей *да*. Последняя употребляется в целом ряде разных контекстов и, по-видимому, многозначна (или даже омонимична). Ср. (i) *Пойдешь с нами? – Да*; (ii) *Да будет свет!*; (iii) *Да что с него спрашивать?*. Неясно, с какой из этих лексем современного слова *да* сближается элемент *да* из *даже*. Кроме того неясно, какова была семантика партикулы *да* на стадии ее склеивания с *же*. Тем не менее можно предположить, что элемент *да* вошел в состав новой частицы *даже* в весьма выветренном виде. При этом элементы *да* и *же* в составе *даже* сплавились, слились в такой степени, что в семантической структуре этой частицы трудно определить, какие ее компоненты «восходят» к *да*, а какие – к *же*.

Частица *даже* – пример очень далеко зашедших процессов выветривания и сплавления «партикул». Не так обстоит дело с союзами *а то* и *а не то*. Недаром в [Грамматике-80] они отнесены к составным, т.е. неоднословным союзам.

Мы попытаемся выяснить, каков вклад в семантику союза каждого из его составных элементов – *а*, *то* и *не*. Иными словами, мы попытаемся подвергнуть эти союзы композициональному анализу.

В современной теоретической семантике композициональный анализ лексики не очень популярен. Другое дело – композициональный анализ высказывания: он применяется в разных семантических теориях. [Богуславский 1996: 9 слл.]. Все же существует научное направление, в котором композициональный подход применяется к «мелким» словам русского языка – оно разрабатывается Д.Пайаром и лингвистами его школы [Пайар 1993; 1998]. Мы в большой степени разделяем идеи Д.Пайара, стараясь при этом принимать во внимание и семантические процессы, постулированные выше.

Подчеркнем, что составные союзы, в частности *а то* и *а не то*, – это не сочетания элементов, а отдельные лексические единицы, поэтому даже установив вклад каждого элемента в семантическую структуру союза, мы не получим толкования союза: эти элементы настолько «сплавились» друг с другом, что значение союза не сводится к сумме значений его компонентов. В результате композиционального анализа лексической единицы мы в лучшем случае можем надеяться получить лишь «костяк» ее значения. Адекватное описание значения лексиче-

082-Uryson_E:_Layout 1 18.05.2008 16:24 Page 532

ской единицы, в частности выявление тех специфических компонентов, благодаря которым данная лексическая единица является именно словом, а не сочетанием слов, обеспечиваются другими способами [Апресян 1974].

Заметим, что современная семантика видит свою основную задачу в том, чтобы адекватно истолковать единицы языка и описать правила их сочетаемости в высказывании (и, разумеется, выявить при этом основные особенности языковой картины мира) [Апресян 1974; 1995]. В конечном счете такое описание может лечь в основу семантического компонента модели типа «СМЫСЛ - ТЕКСТ» С помощью композиционального анализа нашего материала решаются другие задачи.

Во-первых, подвергая данные единицы композициональному анализу, мы надеемся лучше представить логику механизмов десемантизации и сплавления партикул в новую «макромолекулу».

Во-вторых, результат такого анализа может послужить мостом между описанием современной семантики выбранной единицы и ее этимологией. Возьмем, например, союз *если*. С этимологической точки зрения, *если = есть+ли*. Однако с точки зрения синхронного описания современного русского языка, *если* – это семантический примитив, т.е. неразложимая единица. Желательно представлять, каким образом сочетание единиц дало семантически неразложимое слово. Это, по-видимому, невозможно без композиционального подхода к его значению. Другой пример – союз *или*. Этимологически *или = и+ли*. Возникает задача моделировать семантические процессы, в результате которых сочетание данных элементов дало смысл ‘или’. Первый этап такого моделирования – композициональный анализ современного слова.

2. Композициональный анализ союзов **а то *и а* не то**

***2.1. Союзы* а то *и* а не то *как синонимы* или (Гуляли в лесу, а то <а не то, или> шли на озеро).** Союзы *а то* и *а не то* в контекстах типа (1) относятся к разделительным [Грамматика-80]. В.З.Санников пишет об этих союзах, что они синонимичны союзу *или* [Санников 1989: 110]. Ср. пример (1), а также следующие, где *а то* (или *а не то*) заменяется на *или* без существенного изменения смысла:

(5) а. *Гуляли в лесу, а то <а не то> шли на озеро*. – б. *Гуляли в лесу или шли на озеро*.

(6) а. *Иногда засыпала сразу и спала всю ночь, а то <а не то> до рассвета читала романы*. – б. *Иногда засыпала сразу и спала всю ночь или (же) до рассвета читала романы*.

(7) а. *Ходит скучная, ни с кем не разговаривает, а то <а не то> ляжет лицом к стене и плачет*. – б. *Ходит скучная, ни с кем не разговаривает, или ляжет лицом к стене и плачет*.

Однако В.З.Санников приводит и такие примеры, в которых союз *или* нормален, а союз *а то* неудачен или даже невозможен. Ср. примеры [Санников 1989: 110]:

(8а) *Он приедет в пятницу или в субботу*

VS.

(8б) ?*Он приедет в пятницу, а то в субботу*.

(9а) *Катя учится на славянском или на русском отделении*

VS.

(9б) **Катя учится на славянском, а то на русском отделении*.

Правда, в примерах (8)-(9) как будто возможен союз *а не то*. Ср.

(10) *Он приедет в пятницу, а не то в субботу*.

(11) *Катя учится на славянском, а не то на русском отделении*.

Однако (10) и (11), в отличие от их вариантов с *или*, представляют собой не вполне нормативные высказывания – их можно квалифицировать как просторечные. Между тем другие прведенные примеры с *а не то* вполне нейтральны. На наш взгляд, в (10) и (11) единица *а не то* – это просторечный вариант разговорного союза *не то …а не то* с усеченной первой частью. Ср.

(10а) *Он приедет не то в пятницу, а не то в субботу*.

(11а) *Катя учится не то на славянском, а не то на русском отделении*.

Такой союз *а не то* мы не рассматриваем.

Итак, в контекстах типа (8а) – (9а) возможен союз *или*, но не союзы *а то* и *а не то* в рассматриваемых значениях. Этот факт требуется объяснить.

На наш взгляд, союзы *а то* и *а не то* отличаются от *или* семантически, и именно семантика препятствует их свободной взаимозамене в некоторых контекстах.

Союз *или* вводит представление о множестве альтернатив. При этом говорящий (субъект ситуации) знает или уверен, что какая-то альтернатива из множества имеет место в действительности, однако не знает, какая именно. Иными словами, утверждение с *или* предполагает и определенное знание, и незнание ситуации, о которой идет речь. Поэтому союз *или* нормален в высказываниях типа

082-Uryson_E:_Layout 1 18.05.2008 16:24 Page 533

(8г) *Он приедет в пятницу или в субботу* [говорящий знает, что субъект приедет в один из этих дней, но не знает, в какой именно].

(9г) *Катя учится на славянском или на русском отделении* [говорящий знает, что Катя учится на одном из этих отделений, но не знает на каком].

По этой же причине союз *или* возможен как в контексте предиката *знать* и других предикатов знания, так и в контексте *не знать* и вообще предикатов незнания (данные контексты с *или* противопоставлены просодически). Ср.

(10а) *Я знаю <мне известно>, когда он приедет - в пятницу или в субботу*.

VS.

(10б) *Я не знаю <мне неизвестно>, когда он приедет - в пятницу или в субботу*.

Подробнее о союзе *или* см. [Урысон 2004].

Союзы *а то* и *а не то* семантически проще. В рассматриваемом значении (ср. *Гуляли в лесу, а то <а не то> шли на озеро*) они служат для описания чередующихся ситуаций и не указывают ни на знание, ни на незнание чего-либо говорящим.

Предлагаем следующее толкование союзов *а то* и *а не то* в рассматриваемом значении (упрощенно):

(11) *P, а то Q <P, а не то Q>* [*Гуляли в лесу (P), а то <а не то> шли на озеро (Q)*] = ‘в некоторые отрезки времени имеет место ситуация P; в некоторые подобные отрезки времени не имеет место ситуация P; в эти отрезки времени имеет место ситуация Q’.

В случаях (8) – (9) чередующихся ситуаций нет, поэтому в них союзы *а то* и *а не то* в рассматриваемом значении невозможны.

Что касается союза *или*, то он толкуется через компонент ‘возможно’. Ср. толкование союза *или*, предлагаемое В.З.Санниковым: [Санников 1989: 104]:

(14) *P или Q* = ‘возможно P, возможно Q’.

Аналогичным образом толкует союз *или* А. Вежбицкая [Вежбицкая 1996]. Заметим, что в соответствии с нашим подходом компонент ‘возможно’ обязан своим существованием элементу *ли* в составе *или*. Это, однако, тема отдельной работы.

Правда, союз *или* тоже возможен при описании чередующихся ситуаций, ср. (5)-(7). Однако в примерах типа *Гуляли в лесу или шли на озеро* нет прямого указания на то, что описываемые ситуации чередуются. Семантическая структура этого высказывания в первом приближении может быть представлена так:

(5в) ‘в некоторый отрезок времени возможна ситуация P, возможна ситуация Q’.

Мы можем понять, что ситуации P и Q чередуются, только из широкого контекста. Ср.

(5г) *Не скучно было отдыхать в деревне? – Нет, мы гуляли в лесу или шли на озеро, там всюду хорошо* [ситуации чередуются].

(5д) *Как ты думаешь, что они делали вчера вечером? – Не знаю, гуляли в лесу или шли на озеро* [однократные, не чередующиеся ситуации].

Заметим, что ситуация Q, вводимая союзом *а то* или *а не то*, обычно выделена по сравнению с другими ситуациями, перечисляемыми в данной сочинительной конструкции. Подробнее об этом см. [Израэли 2000; Санников 1989: 182]. Можно думать, что отчасти в связи с этим *а то* образует цельное сочетание *а то всё*. Ср. примеры, приводимые в книге [Санников 1989: 182]:

(12) *Перелетает только через реки и моря, а то все пешком* (А.П.Чехов).

(13) *Лишь изредка подобие леса, какой-нибудь Заказ, Дубровка, а то все поля, поля, беспредельный океан хлебов* (И.Бунин).

Эту тонкую семантику союзов *а то* и *а не то* мы пока не рассматриваем.

Теперь перейдем к нашему основному вопросу.

***2.2.Элемент* то *в составе разделительных союзов* а то *и* а не то.** Абстрагируемся от частицы *-то*, представленной в случаях типа *Он-то придет, Дом-то у них большой, удобный*, а также от *то* как второй части союзов типа *если ...то*, *когда ... то* и т.п. Во многих других случаях слово *то* – это форма среднего рода им.-вин. падежа местоимения *тот*. В своем центральном значении это указательное местоимение, т.е. дейктическое слово. Ср. *Да он не в этот дом вошел, а вон в тот* (И.А.Гончаров, [МАС]); *- Вон взгляни, например, на то дерево. Куда оно годится?* (Ю.Лаптев, [МАС]).

Дейктические слова обычно имеют анафорическое значение. Ср.

(15) *Сядем под это дерево* [*этот* – дейктическое слово]

VS.

Он сейчас переводит «Евгения Онегина» и берет эту книгу во все поездки [*этот* – анафорическое слово].

(16) *Тебе очень пойдет такое платье* [с указательным жестом, напрввленным на конкретный предмет, *такой* – дейктическое слово]

082-Uryson_E:_Layout 1 18.05.2008 16:24 Page 534

VS.

Такие случаи теперь не редкость [*такой* – анафорическое слово].

(17) *Делай так* [адресату показывают как; *так* – указательное слово]

VS.

Так бывает со всеми новичками [*так* – анафорическое слово].

Слово *тот* тоже имеет анафорические значения. Правда, многие высказывания с анафорическим *тот* сейчас воспринимаются как устаревшие. Ср. *Людей без гордости и сердца презирает, / А сам игрушка тех людей* (М.Ю.Лермонтов, [МАС]) [сейчас было бы естественнее: ... *этих людей*]. – *Это было в 18... таком-то году. В то время я только что получил место и ехал с товарищем на прииск* (Короленко, МАС). В приведенных примерах слово *тот* выступает как местоимение-прилагательное, согласованное с определяемым существительным. Оно отсылает к объекту, упомянутому ранее.

Однако данное местоимение может выступать и как существительное среднего рода *то*. Ср.

(18) *Я убил / Супруга твоего; и не жалею / О том – и нет раскаяния во мне* (А.С.Пушкин, Каменный гость).

(19) *Матушка занималась хозяйством; меня ничему не учили, а я тому и рада была* (Ф.М.Достоевский, [МАС]).

В этих примерах *то* выступает как чистое анафорическое слово, причем оно отсылает не к объекту (как местоимение-прилагательное *тот*), а к ситуации.

В следующих примерах семантика анафорического *то* несколько богаче. Ср.

(20) *Каждый почти вечер видно зарево далеких пожаров (P): то турки жгут болгарские деревни (Q)* (Гаршин, МАС).

(22) *Изредка от края до края мола перекатывался заунывный шум (P) – то спросонок разбивалась о камень волна (Q)* (К.Паустовский, МАС).

Здесь *то*, как и в предыдущих примерах, отсылает к ранее названной ситуации (P). Однако в (20)-(22) *то* не только выражает анафору, но еще и идентифицирует ранее названную ситуацию P с ситуацией Q, обозначаемой следующим предложением. При этом ситуация P может только подразумеваться. Ср.

(23) *То не ветер ветку клонит, / Не дубравушка шумит, / То мое сердечко стонет, / Как осенний лист дрожит* (Стромилов, МАС).

Употребление *то* в примерах (18)-(23) явно отмечено стилистически: в современном русском языке в подобных контекстах употребляется слово *это*. И *то*, и *тот* как анафорические показатели уходят из языка.

Все же и в современном языке существуют нейтральные контексты с анафорическим *то*. Правда, эти контексты обладают определенной синтаксической спецификой – *то* выполняет в них функцию т.н. указательного слова в главном предложении. Ср.

(24) *Что случилось, того уж не вернешь* (А.П.Чехов, МАС).

(25) *Мы не поверили тому, что нам рассказали*.

Как бы то ни было, местоимение-существительное *то* является анафорическим словом как минимум в трех разных типах контекстов, ср. (18)-(19), (20)-(23), (24)-(25). Естественно считать, что перед нами три очень близких лексемы анафорического *то*. Все эти лексемы отсылают к ситуации. Обозначение этой ситуации часто находится в предтексте, ср. (18)-(24). Однако это не обязательно. Ср. (25), где *то* отсылает к последующему фрагменту высказывания. Тем самым, *то* как отсылочное слово может отсылать и к предшествующему фрагменту текста (анафора в узком смысле слова), и к последующему фрагменту (катафора).

Элемент *то* в союзах *а то* и *а не то* напрямую не соотносится ни с одной лексемой местоимения-существительного *то* (и ни с одной лексемой местоимения *тот* вообще). Тем не менее данный элемент в составе рассматриваемых союзов относительно мало выветрен и сохраняет свое анафорическое значение. К чему же он отсылает?

Высказывания вида *P, а не то Q* естественно сближаются с микротекстами типа P^1. *Не P – Q*. Ср.

(26) [*P, а не то Q*] *Ходили за грибами (P), а не то брали лодку и плавали по озеру (Q)*. VS.

[P^1. *Не P – Q*] *Ходили за грибами* (P^1). *Не шли за грибами (P) – брали лодку и плавали по озеру (Q)*.

Здесь фрагмент P – это, по существу, повторение предшествующего фрагмента P^1 (это повторение с точностью до видо-временной семантики, референциальных статусов и т.п.). Следовательно, P можно заменить анафорическим словом. Язык выбрал для этого слово *то*. В результате получается цепочка *не то*.

(26а) *Ходили за грибами* (P^1), *не то* [‘не P^1’; *то* отсылает к P^1] *брали лодку и плавали по озеру (Q)*.

В (26а) нет эксплицитного обозначения ситуации P – вместо него в высказывании анафорический элемент *то*, отсылающий к обозначению ситуации в предтексте.

Микротексты типа (26) естественны с союзом *а*. Ср.

(27) *Ходили за грибами* (P^1), *а не шли за грибами (P) – брали лодку и плавали по озеру (Q)*.

082-Uryson_E:_Layout 1 18.05.2008 16:24 Page 535

Союз *а* естественно предваряет и цепочку *не то*. В результате имеем *а не то*. Ср.

(27а) *Ходили за грибами (P[1]), а не то* [‘а не P[1]’; *то* отсылает к P[1]] *брали лодку и плавали по озеру (Q)*.

Аналогичный пример:

(28) *Обычно спал хорошо (P[1]). А не спалось (P) – читал до рассвета романы (Q)*.

(29)*Обычно спал хорошо (P[1]), а не то* [*то* отсылает к P[1]] – *читал до рассвета романы (Q)*.

В соответствии с этим анализом, в союзе *а не то* [ср. *Гуляли в лесу (P), а не то шли на озеро (Q)*] элемент *то* отсылает к ситуации P, т.е. к предыдущему фрагменту текста (анафора). С некоторой долей условности, можно сказать, что антецедентом *то* в данном случае является P.

Союз *а то* [ср. *Гуляли в лесу (P), а то шли на озеро (Q)*] отличается от *а не то* следующим: здесь элемент *то* отсылает к ситуации Q, т.е. к линейно последующему фрагменту текста (катафора).

Иными словами, союзы *а то* и *а не то* в рассматриваемом разделительном значении различаются направлением отсылки. В союзе *а не то* элемент *то* отсылает к предтексту. В союзе *а то* этот же элемент катафоричен.

Перейдем к следующей лексеме союзов *а то* и *а не то*.

2.3. Контексты угрозы **(Отдай машинку, а то <а не то> маме скажу).** Данный тип контекстов иллюстрируется примерами типа

(30) *Придется дать кролику капусты, а то <а не то> он убежит*.

(31) *Отдай машинку, а то <а не то> маме пожалуюсь*.

(32) *Держу вас только из уважения к вашему почтенному батюшке, а то бы вы у меня давно полетели со службы* (А.П.Чехов, [Санников 1989]) [здесь возможен и союз *а не то*].

Возьмем пример (32а) с *а не то*:

(32а) *Держу вас только из уважения к вашему почтенному батюшке, а не то бы вы у меня давно полетели со службы*.

Здесь первый сочиненный компонент выражает пресуппозицию ‘я уважаю вашего почтенного батюшку’. Кроме того, в этом высказывании выражен следующий смысл: ‘если бы я не уважал вашего батюшку, вы бы у меня давно полетели со службы’ (наличие условного смысла в подобных контекстах с *а то* и *а не то* отмечается всеми исследователями). Следовательно, в семантической структуре примера (30) есть следующий фрагмент:

(32б) (i) [пресуппозиция] ‘я уважаю вашего почтенного батюшку’.

(ii) ‘если бы я не уважал вашего батюшку, вы бы у меня давно полетели со службы’.

В более эксплицитном виде смысл (32б) выражен в следующем примере:

(32в) *Держу вас только из уважения к вашему почтенному батюшке. Если бы я его не уважал (P), вы бы у меня давно полетели со службы (Q)*.

Условное отношение *Если P, Q* может быть выражено и бессоюзно – с помощью порядка слов и интонации. Ср.

(32г) *Держу вас только из уважения к вашему почтенному батюшке. Не не уважал бы я его (P), вы бы у меня давно полетели со службы (Q)*.

И в (32в), и в (32г) фрагмент P повторяет пресуппозицию ‘я уважаю вашего почтенного батюшку’. Разумеется, это повторение с точностью до некоторых значений, например значения наклонения. Тем не менее, P можно заменить анафорическим словом – анафорическая отсылка не предполагает буквального совпадения смыслов (см. [Падучева 1990]). В качестве такого анафорического слова используется *то*. В результате получается *не то*, а с предшествующим *а* – союз *а не то*. Ср.

(32г) *Держу вас только из уважения к вашему почтенному батюшке* [пресуппозиция: ‘я уважаю вашего почтенного батюшку’ (P[1])], *(а) не то* [‘а не P[1]’; *то* отсылает к P[1]], *вы бы у меня давно полетели со службы (Q)*.

Теперь возьмем этот же контекст с *а то*, ср.

(32) *Держу вас только из уважения к вашему почтенному батюшке, а то бы вы у меня давно полетели со службы*.

Здесь выражена та же пресуппозиция (32б). Но это – не вся пресуппозиция фраз (32) и (32а). В частности,

(32б) (ii) ‘если бы я не уважал вашего батюшку, вы бы у меня давно полетели со службы’

содержит следующую пресуппозицию:

(32б) (iii) ‘я мог бы не уважать вашего почтенного батюшку’.

Элемент *то* союза *а то* отсылает именно к этому фрагменту семантической структуры предтекста. Иными словами, семантическая структура высказывания с *а не то* и с *а то* строится из одних и тех же элементов. Но отрицание ‘не’ в первом случае является составной частью союза *а не то*, а в данном случае входит в состав фрагмента, к которому отсылает элемент *то*.

Аналогичным образом устроен пример (31) – здесь элемент *то* союза *а не то* тоже отсылает к смыслу, выраженному в первом фрагменте высказывания. Действительно, императив P выражает желание говорящего, чтобы имела место ситуация P[1] ‘ты отдаешь мне машинку’. Семантическая структура высказывания (31) содер-

082-Uryson_E:_Layout 1 18.05.2008 16:24 Page 536

жит следующий фрагмент:

(31а) (i) [цель речевого акта] ‘ты отдаешь мне машинку’.

(ii) ‘если ты не отдашь мне машинку, я пожалуюсь маме’.

В более эксплицитном виде этот смысл выражен в примерах:

(31б) *Отдай машинку. Если не отдашь (Р), пожалуюсь маме*.

(31в) *Отдай машинку. Не отдашь (Р), пожалуюсь маме*.

Фрагмент Р здесь повторяет (с точностью до значения наклонения и, возможно, некоторых других) один из семантических компонентов первой части ‘ты отдаешь мне машинку’. Следовательно, Р можно заменить анафорическим словом. В качестве такого анафорического слова используется *то*. В результате получается *не то*, а с предшествующим *а* – союз *а не то*. Ср.

(31г) *Отдай машинку* [‘(я хочу, чтобы) ты отдал мне машинку’ (P^1)], *а не то* [‘а не P^1’; *то* отсылает к P^1] *маме пожалуюсь*.

Возьмем теперь тот же контекст с *а то*. Ср.

(31д) *Отдай машинку, а то маме пожалуюсь*.

Компонент

(31а) (i) [цель речевого акта] ‘ты отдаешь мне машинку’

в свою очередь имеет следующую пресуппозицию:

(31а) (iii) ‘возможно, ты не отдаешь мне машинку’.

В (31д) элемент *то* союза *а то* отсылает именно к этому компоненту семантической структуры предтекста. Этот компонент содержит отрицание ‘не’, поэтому никакого отрицания при анафорическом элементе не требуется.. (В союзе *а не то* элемент *то* отсылает к компоненту, который нужно подвергнуть отрицанию, и это отрицание выражено частицей *не* при элементе *то*.)

Аналогичным образом устроен пример (30). Здесь первый сочиненный компонент выражает импликатуру ‘мы дадим кролику капусту’. Кроме того, в этом высказывании выражен следующий смысл: ‘если не дадим кролику капусту, он убежит’. Следовательно, в семантической структуре примера (30) есть следующий фрагмент:

(30а) (i) [следствие] ‘мы дадим кролику капусту’.

(ii) ‘если не дадим кролику капусты, он убежит’.

В более эксплицитном виде смысл (30а) выражен в следующем примере:

(30б) *Придется дать кролику капусты. Если не дать ему капусты (Р), он убежит (Q)*.

Условное отношение *Если Р, Q* может быть выражено и бессоюзно – с помощью порядка слов и интонации. Ср.

(30в) *Придется дать кролику капусты. Не дать <не дашь> ему капусты (Р) – он убежит (Q)*.

И в (30б), и в (30в) фрагмент Р включает в себя импликатуру ‘мы дадим кролику капусту’, выраженную в предшествующем фрагменте. Поэтому Р можно заменить анафорическим словом. В качестве такого анафорического слова используется *то*. В результате получается *не то*, а с предшествующим *а* – союз *а не то*. Ср.

(30г) *Придется дать кролику капусты* [импликатура ‘мы дадим кролику капусту’ (P^1)], *(а) не то* [‘а не P^1’; *то* отсылает к P^1], *он убежит (Q)*.

В том же контексте с *а то* элемент *то* отсылает к пресуппозиции

(32) (iii) ‘мы не даем кролику капусты’.

Отрицание ‘не’ содержится в данном фрагменте, поэтому отсылка к нему оформляется элементом *то* без частицы *не*.

Итак, *а не то* в данном значении отсылает к фрагменту семантической структуры, который нужно подвергнуть отрицанию. Это отрицание выражается элементом *не* союза. Союз *а то* отсылает к фрагменту семантической структуры, котрый сам уже содержит отрицание, поэтому в союзе отрицание отсутствует.

Заметим, что оба союза *а то* и *а не то* допустимы не только в контексте угрозы, но и других модальных когнтекстах. Ср. пример из книги В.З.Санникова:

(33а) *Я вот смеяться могу, и смеюсь, а не то бы, верно, плакал* [элемент *то* отсылает к смыслу ‘я могу смеяться’; *не то* - ‘не мог бы смеяться’].

VS.

(33б) *Я вот смеяться могу, и смеюсь, а то бы, верно, плакал* [в первой части выражена пресуппозиция ‘возможно, я не мог бы смеяться’; элемент *то* отсылает к ней].

Заметим, что союз *а то* в контекстах угрозы (и в других модальных контекстах) допускает и более простую интерпретацию. Возможно, элемент *то* в этом случае катафоричен и просто отсылает к последующей ситуации. Тогда рассматриваемые примеры отчасти сближаются с бессоюзными высказываниями, оформленными специфической просодией. Ср.

082-Uryson_E:_Layout 1 18.05.2008 16:24 Page 537

(33) *Придется дать кролику капусту (P) – убежит (Q)*

VS.

Придется дать кролику капусту (P) – а то ['а Q'; элемент *то* отсылает к последующемй фрагменту Q] *убежит (Q)*

(34) *Отдай машинку, маме пожалуюсь*.

(32) *Держу вас только из уважения к вашему почтенному батюшке – вы бы у меня давно полетели со службы*.

Несколько слов об элементе *а* союзов *а то* и *а не то*. Семантически он сближается с союзом «*а* поворота повествования». Этот союз представлен, например, в следующих контекстах: *Стою, жду автобуса. А мороз двадцать градусов*; *Я завтра уезжаю. А поезд у меня днем, так что вещи складывать буду сегодня*; *Катя готовилась к экзамену по химии, а это был единственный предмет, который она знала плохо*. Подробнее о союзе «*а* поворота повествования» см. [Урысон 2002; 2006].

***2.4. Союз* а то, *не заменимый на* а не то.** Союз *а то* имеет еще одну (разговорную) лексему. Она вводит указание причины [Белошапкова 1970; Санников 1989]. Ср.

(33) *Пойду, а то уже поздно* ['пойду, потому что уже поздно'].

Шестьдесят копеек оставила себе, шестьдесят отложила Лидии Афанасьевне, а то ей не на что было возвращаться (Ю.Трифонов, пример В.А.Белошапковой).

Элемент *то* вводит здесь обозначение ситуации-причины. Вопрос о причине в разговорной речи вводится словм *что*, ср. *Что он такой расстроенный?, Что ты так торопишься?* и т.п. [Арутюнова 1980]. Если в вопросе причина обозначается словом *что*, то в утверждении она естественно маркируется словом *то*. Элемент *то* данной лексемы *а то* отсылает к последующему фрагменту (катафора). Союз *а не то* в таких случаях невозможен, ср. **Пойду, а не то уже поздно*. Причина в том, что в семантической структуре подобного высказывания нет компонента, к которому мог бы отсылать элемент *то* с отрицанием.

Союз *а то* употребляется так же для введения ситуации, сменившей ситуацию, описанную в предтексте. Ср.

Теперь ей лучше, а то все болела.

Несколько ночей прошло, пока научились <...> находить свой овраг, а то плутали (Ю.Трифонов, [Санников 1989: 182]).

Элемент *то* здесь катафоричен. Союз *а не то* в подобных случаях не употребляется.

Аналогичным образом устроено употребление *а то* в качестве отдельной реплики, ср. *Пойдешь на каток? – А то!* Здесь *А то!* – это, по-видимому, «сокращенный» вариант реплики *А то нет!* Элемент *то* и в этом случае отсылает к последующему фрагменту (*нет*).

В целом, *то* в составе союза *а не то* отсылает к предтексту. В составе союза *а то* этот же элемент оьычно отсылает к последующему тексту.

Если это так, то союзы *а то* и *а не то* различаются направлением отсылки, выражаемой элементом *то*. В составе союза *а то* этот же элемент отсылает к последующему фрагменту текста. В составе *а не то* он отсылает к некоторой ранее названной или подразумеваемой ситуации.

3.Заключение

Мы попытались выявить семантический вклад элементов *а*, *то не* в семантику некоторых лексем союзов *а то* и *а не то*. Для этой цели мы возводили высказывание, содержащее данный союз, к некоторому другому высказыванию. Это второе высказывание обладает прозрачной синтаксической структурой, в которой ясно выделяется фрагмент, заменяемый анафорическим словом *то*.

В соответствии с нашим описанием элемент *то* в союзе *а не то* отсылает к одному фрагменту семантической структуры высказывания, а в союзе *а то* - к другому. Поэтому *а то* и *а не то* не различаются как другие пары типа 'X' – 'не X'. Почему же они синонимичны?

Оба союза описывают некоторое положение дел, состоящее из двух ситуаций: P и Q. Эти ситуации исключают друг друга: речь идет о том, что не имеет место P и имеет место Q. Союз *а не то*, отсылая к P, указывает, что P не имеет место. Союз *а то* отсылает к Q и указывает, что Q имеет место. Тем самым, синонимия союзов обусловлена контекстом анафорического элемента *то*.

Интегральный подход к описанию языка, развиваемый в московской семантической школе [Апресян 1995], заключается в согласовании грамматики и словаря данного языка. Можно также говорить о согласовании синхронного описания языковых единиц с их этимологией и историей. Такое согласование предполагает не только использование единого метаязыка описания, но и применение методов декомпозиции, отличных от тех, что выработаны для истолкования языковой единицы (последние описаны в Ю.Д.Апресяном в книге [Апресян

082-Uryson_E:_Layout 1 18.05.2008 16:24 Page 538

1974]). Мы попытались показать это на примере союзов *а то* и *а не то*. Композициональный анализ других союзов и частиц с элементом *то* может подтвердить, уточнить или опровергнуть предлагаемую интерпретацию выбранного материала.

Список литературы

1. Апресян Ю.Д. Лексическая семантика. М.: «Наука», 1974.

2. Апресян Ю.Д. Интегральное описание языка и системная лексикография // Избранные труды. Т. II. М.: «Языки русской культуры», 1995.

3. Арутюнова Н.Д. Оценка. Событие. Факт. М.: «Наука», 1980.

4. Белошапкова В.А. Предложения альтернативной мотивации в современном русском языке // Исследования по современному русскому языку. М.: Издательство Московского университета, 1970. С. 13 – 29.

5. Богуславский И.М. Сфера действия лексических единиц. М.: «Языки русской культуры», 1996.

6. Вежбицкая А. The semantics of logical concepts // Московский лингвистический журнал. Т. 2. М. 1996.

7. Грамматика-80 Русская грамматика. Т. I-II. М.: «Наука», 1980.

8. Israeli A. The meaning and polysemy of the alternative conjunction a to. Manuscript.

9. Inkova-Manzotti O. Encore sur la conjonction Russe A TO . Manuscript.

10. Колосова Т.А. О сигналах неразвернутости некоторых имплицитных сложных предложений // Синтаксис предложения. Калинин. 1980.

11. МАС – Толковый словарь русского языка в четырех томах. Т. 1 – 4. М. 1985 – 1990.

12. Николаева Т.М. Функции частиц в высказывании: на материале славянских языков. М.: «Наука», 1985.

13. Николаева Т.М. Пространство славянских партикул // Славянское языкознание. XIII Международный съезд славистов. Любляна, 2003 г. Доклады российской делегации. М., 2003.

14. Падучева Е.В. Анафорическое отношение // Лингвистический энциклопедический словарь. М.: «Наука», 1990.

15. Пайар Д. Путеводитель по дискурсивным словам русского языка / Научный руководитель проекта Д.Пайар. М., 1993.

16. Пайар Д. Дискурсивные слова русского языка / Научный руководитель проекта Д.Пайар. М., 1998.

17. Подлесская В.И. ИНАЧЕ, А ТО, А НЕ ТО: резумптивные союзы как способ выражения отрицательного условия // Сложное предложение: традиционные вопросы теории и описания и новые аспекты его изучения. Вып. 1. М., 2000.

18. Санников В.З. Русские сочинительные контсрукции. М.: «Наука»,1989.

19. Собинникова В.И. Сложные предложения с союзом а то в русских и украинских говорах // Материалы по русско-славянскому языкознанию. Т. 4. Воронеж. 1969.

20. Урысон Е.В. Союз А как сигнал поворота повествования // Логический анализ языка. Семантика начала и конца / Отв. ред. чл.-корр. РАН Н.Д. Арутюнова. М.: «Индрик», 2002.

21. Урысон Е.В. – Словарная статья «ИЛИ (… ИЛИ), ЛИБО (… ЛИБО), НЕ ТО … НЕ ТО, ТО ЛИ … ТО ЛИ» // Новый объяснительный словарь синонимов русского языка. 2-е изд., исправл. и дополн. М.: «Языки славянской культуры», 2004.

22. Урысон Е.В. Подсистема русских сочинительных союзов И, А, НО // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции «ДИАЛОГ 2006». М., 2006.

083-Ury:_Layout 1 18.05.2008 16:25 Page 539

АВТОМАТИЧЕСКОЕ РАЗБИЕНИЕ ТЕКСТА НА ПРЕДЛОЖЕНИЯ ДЛЯ РУССКОГО ЯЗЫКА

DETECTING SENTENCE BOUNDARIES IN RUSSIAN

Урюпина О. *(uryupina@gmail.com)*
Институт языкознания РАН,
«Ашманов и Партнеры»

В данной работе предлагается статистический алгоритм сегментации текста на предложения на материале русского языка. Алгоритм основан на контекстах знаков препинания и не требует предварительного синтаксического анализа. Сравнение с эвристическими методами показывает, что статистический подход позволяет существенно улучшить качество сегментации.

1. Введение

Большинство систем автоматической обработки языка ставят своей задачей анализ текстов, заранее разбитых на предложения. Например, парсеры определяют синтаксическую структуру предложения, системы автоматического реферирования выделяют из документа наиболее значимые предложения и так далее. В то же время, языковые данные доступны нам чаще всего в виде текстов, размеченных на абзацы, главы и другие более крупные единицы. Поэтому для их эффективного автоматического анализа необходимы соответствующие алгоритмы сегментации. К сожалению, большинство систем используют упрощенные эвристические методы разбиения на предложения. Нам не удалось найти подробных теоретических исследований данной задачи или описаний работающих алгоритмов для русского языка, за исключением очень кратких (см., например, [6]). Задача разбиения текста на предложения для английского языка описывается, в частности, в [7] и [8].

Рассмотрим простой пример, показывающий, как неправильное разбиение на предложения может породить целый ряд проблем на разных уровнях анализа документа:

(1) Но ведь Маша знает А. Б. Иванова много лет и никогда про него ничего плохого не слышала!!

Если мы ошибочно выделим здесь четыре предложения («Но ведь Маша знает А.», «Б.», «Иванова много лет и никогда про него ничего плохого не слышала!» и «!»), то дальнейшая автоматическая обработка окажется бессмысленной. Так, парсеры либо не смогут разобрать предпоследнее предложение, либо сделают это неверно. Интерпретация местоимения «него» окажется затруднительной – скорее всего, будет принято решение, что «него» кореферентно с «Б». Попытки использовать этот фрагмент в экспертной системе приведут к тому, что на вопрос «Что знает Маша?» будет выдан ответ «А». Наконец, модуль автоматического реферирования может включить одно из четырех ошибочно выделенных предложений в аннотацию документа, что приведет к потере качества.

В данной работе обсуждается алгоритм разбиения текста на предложения. Рассматриваются две связанные задачи:

1) определение, является ли терминальный знак препинания (здесь и далее под терминальными знаками мы будем понимать точку, восклицательный и вопросительный знаки) границей предложения в данном контексте,

2) определение всех границ предложений в документе.

Мы предлагаем статистический алгоритм, который можно легко адаптировать для анализа текстов разных типов (например, книг vs. веб-страниц). Важной особенностью нашего алгоритма является то, что он не опирается на синтаксический анализ. Это дает нам, во-первых, существенный выигрыш в скорости и, во-вторых, возможность анализировать практически любые тексты, независимо от их синтаксической грамотности.

В разделе 2 мы приводим примеры из Национального Корпуса Русского Языка [1], иллюстрирующие сложность задачи. В разделе 3 описываются лингвистические признаки, использованные для обучения и распознавания. В разделе 4 обсуждаются результаты экспериментов: мы сравниваем эффективность нашего алгоритма и нескольких интуитивных стратегий выделения предложений (например, «если после терминального знака идет слово с большой буквы, то это новое предложение»).

083-Ury:_Layout 1 18.05.2008 16:25 Page 540

2. Примеры

В самом первом приближении можно считать, что предложение всегда начинается со слова с большой буквы и заканчивается терминальным знаком препинания. Однако и теоретические исследования (см., например, [2]), и корпусные данные говорят о том, что для наиболее точного и полного выделения предложений необходимо учитывать целый ряд дополнительных факторов.

Довольно часто точка является разделителем не предложения, а других единиц. Например, точка используется в URL веб-страниц (2, здесь и далее все примеры взяты из Национального Корпуса Русского Языка) и для обозначения даты или времени (3):

(2) Конечно, в рамках газетной статьи невозможно сделать обзор сотни докладов, поэтому мы рекомендуем посетить Интернет-страницу конференции http://www.ict.nsc.ru/ws/mol2000/, на которой размещены программа мероприятия и тезисы докладов.

(3) В 11.45 дали слово Кудрину , но он всё не шёл.

Точка используется как знак сокращения:

(4) выполнение 12 тепловозам усиленного ТР-1 с применением средств диагностики вместо ТР-2 дало экономический эффект 221,8 тыс. руб.

Стоит обратить внимание, что после сокращения в конце предложения ставится одна точка, а не две («руб.», а не «руб..»). Это затрудняет анализ: даже если мы знаем, что перед точкой сокращение, мы не можем с уверенностью сказать, что перед нами середина предложения.

Точка может быть элементом форматирования:

(5) Параметры системы питания:
линейное напряжение на проводах 81 ... 83 , В 220
фазное напряжение на проводах 81 ... 83 по отношению
к проводу 84 (нулю), В 127
частота переменного тока, Гц 50
выпрямленное напряжение на проводах, В
15—30 110
44—30 50

Нередко точка в середине предложения просто опечатка:

(6) Режиссёр Михаил. Бычков поставил в Таллине притчу о любви к невозможному и о презрении к реальности

Вопросительный и восклицательный знаки также могут употребляться в середине предложения. Чаще всего ими завершаются фрагменты в скобках (7) или кавычках (8):

(7) Дело в том, что по всяким планам «пятилеток» и заданиям ЦК советский военный комплекс создавал ядерное оружие с запасом на пять и более(!) ядерных войн.

(8) Пролетели «Тише!» Виктора Косаковского и «Фрески» Александра Гутмана.

Часто предложения заканчиваются многоточием (9) или комбинацией вопросительных и/или восклицательных знаков (10). Графически это несколько терминальных знаков, но только последний из них является концом предложения:

(9) Не в лесе и не в медицине дело...

(10) Только вот ради чего ?!

Многоточие может использоваться для передачи паузы или пропуска части текста. Две или три точки могут обозначать интервал между числами. Ни одна из точек не является концом предложения в подобных случаях:

(11) В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал — 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала — 1225 тыс. км).

Определенную проблему представляют скобки и кавычки:

(12) В случае нарушений выносится письменное предупреждение: «В вашей деятельности допускаются вот такие недочёты ...»

В подобных случаях система автоматического разбиения на предложения должна понять, что границей является не терминальный символ, а следующие за ним скобка или кавычки. К сожалению, во многих документах не соблюдаются правила постановки знаков препинания в предложениях со скобками и кавычками, что затрудняет анализ. Далеко не во всех текстах на русском языке используются кавычки-елочки. Анализировать же кавычки-лапки значительно сложнее.

083-Ury:_Layout 1 18.05.2008 16:25 Page 541

Дополнительные проблемы возникают при работе с веб-данными. Так, первое слово в предложении может начинаться с маленькой буквы. Точка может отсутствовать, особенно в конце абзаца. Вопросительный знак может появиться из-за проблем с кодировкой (например, в него могут превратиться кавычки-елочки). Наконец, пробел после терминального знака ставится далеко не регулярно.

Как показывают все эти примеры, сегментация текста на предложения – это довольно сложная задача, требующая учета самых разных факторов. Ниже мы предлагаем статистический алгоритм, основанный на контекстах терминальных знаков препинания.

3. Методология

Наш алгоритм работает следующим образом: сначала из документа извлекаются *контексты* потенциальных границ предложений, потом они описываются в виде *векторов признаков*, которыми оперируют системы автоматического обучения. В нашем первом эксперименте потенциальными границами предложений являются терминальные знаки, во втором – вообще вся пунктуация. Мы не рассматриваем предложения, не заканчивающиеся никаким знаком препинания.[1]

В контекст границы мы включаем четыре элемента: сам знак препинания (punct), ближайшее псевдослово слева (left), ближайшее псевдослово справа (right) и ближайшее собственно слово справа (wright). Под псевдословом здесь и далее понимается любая последовательность символов, не включающая пробел или конец абзаца. Под словом – псевдослово, содержащее хотя бы одну букву или цифру. Также запоминается количество псевдослов слева (dleft) и справа (dright) до ближайшей потенциальной границы или конца абзаца. В Таблице 1 приводятся контексты для всех потенциальных границ в примере (1).

punct	.	.	!	!
left	А	Б	слышала	!
right	Б	Иванова	!	-отсутствует-
wright	Б	Иванова	-отсутствует-	-отсутствует-
dleft	4	1	11	0
dright	1	11	0	0

Таблица 1. Контексты для всех потенциальных границ примера (1) в предположении, что в абзаце нет других предложений.

Составленные таким образом контексты являются основой для векторов признаков. Как показывают примеры из раздела 2, для эффективного разбиения на предложения необходимо учитывать целый ряд факторов: сокращения, вид псевдослов (пунктуация, цифры и т.д.). Эта информация была добавлена в виде дополнительных признаков. В Таблице 2 приведены полные вектора признаков для (1).

Прежде всего, мы составили словарь сокращений. Для этого из размеченной части Национального Корпуса Русского Языка (около 6 миллионов слов) были извлечены все триграммы вида «псевдослово точка слово_со_строчной_буквы». Сокращения, встретившиеся только один раз, были отброшены. Данный словарь был составлен полностью автоматически и не подвергался никакой пост-обработке. Отметим, что ручная разметка корпусных данных никак не использовалась. Таким образом, для составления подобного словаря нужна только большая коллекция текстов. Каждый контекст проверяется на сокращения: если псевдослова left или right нашлись в словаре, то контекст получает дополнительные признаки abbleft и abbright соответственно.

Мы также провели классификацию псевдослов. В отдельные группы были выделены пунктуация и числа. Остальные псевдослова были разбиты на классы в зависимости от используемых символов (кириллица, латиница, кириллица+латиница, кириллица+цифры и так далее) и регистра первого символа (строчная буква, прописная буква, цифра, пунктуация). Класс псевдослов описывается признаками cleft и cright.

Ближайшее собственно слово справа, wright, описывается одним дополнительным признаком – регистр первого символа (cwright).

Наконец, мы ввели дополнительные признаки isfirst и islast для обозначения контекстов, приходящихся на самое начало или конец абзаца.

[1] В наших данных не встретилось предложений, не заканчивающихся знаками препинания, но при этом не приходящихся на конец абзаца или заголовка. Тем не менее, наш алгоритм (после необходимого переобучения и тестирования) применим и для распознавания подобных случаев.

083-Ury:_Layout 1 18.05.2008 16:25 Page 542

punct	.	.	!	!
left	А	Б	слышала	!
right	Б	Иванова	!	-отсутствует-
wright	Б	Иванова	-отсутствует-	-отсутствует-
dleft	4	1	11	0
dright	1	11	0	0
abbleft	1	1	0	0
abbright	1	0	0	0
cleft	-кириллица- -прописная-	-кириллица- -прописная-	-кириллица- -строчная-	
cright	-кириллица- -прописная-	-кириллица- -прописная-	-пунктуация-	-отсутствует-
cwright	-прописная-	-прописная-	-отсутствует-	-отсутствует-
isfirst	0	0	0	0
islast	0	0	0	1

Таблица 2. Вектора признаков для всех потенциальных границ примера (1) в предположении, что в абзаце нет других предложений

Вектора признаков используются для обучения и распознавания, как описывается в следующем разделе.

4. Эксперименты

Ниже описываются языковые данные и системы машинного обучения, использованные для проверки эффективности нашего метода.

Мы проверили вручную 33 документа из Национального Корпуса Русского Языка и исправили все неточности в определении границ предложений. Документы включали в себя газетные и журнальные статьи общего (политика, культура и так далее) и технического (ремонт локомотивов) содержания. Было получено 1639 предложений, 1414 из них заканчивались терминальным знаком. Из этих предложений было выделено 5230 контекстов: 2048 с терминальным знаком и 3182 с другой пунктуацией. Контексты были преобразованы в вектора признаков в соответствии с описанием, изложенным выше. Мы отобрали случайным образом 1000 векторов для тестирования, остальные были зарезервированы для обучения. Были протестировали три алгоритма машинного обучения: C4.5 [3], Ripper [4] и SVM-light [5].

Для оценки эффективности нашего алгоритма мы разработали несколько упрощенных эвристических подходов к определению границ предложений. Прежде всего, конец абзаца всегда считается концом предложения. Наша первая эвристика, termpunct, классифицирует каждый терминальный знак как конец предложения. При анализе примера (1) такой метод приведет к выделению четырех предложений.

Вторая эвристика, termpunct_cap, аналогична первой, но запрещает предложения, не начинающиеся с заглавной буквы. При анализе примера (1) такой метод приведет к выделению трех предложений: «Но ведь Маша знает А.», «Б.», «Иванова много лет и никогда про него ничего плохого не слышала!!».

Наконец, третья эвристика, advanced, дополнительно запрещает предложения, заканчивающиеся сокращением и точкой. При анализе примера (1) такой метод приведет к выделению одного предложения. Именно такой эвристикой или ее незначительными модификациями, насколько нам известно, руководствуются большинство систем автоматического анализа текста для подготовки исходных данных.

В таблице 3 приводятся результаты наших экспериментов. В первом эксперименте рассматриваются только контексты, содержащие терминальные знаки препинания. Мы выделили их в отдельную группу, поскольку большинство предложений заканчиваются именно одним из терминальных знаков. Кроме того, многие алгоритмы, описанные в литературе (см., например, [7]), не ставят своей целью анализ других знаков препинания. Во втором эксперименте рассматриваются все контексты.

Как показывают результаты, с помощью простейшей эвристики termpunct невозможно добиться удовлетворительной точности распознавания: каждая третья поставленная граница будет ошибочной. Остальные эвристики дают более приемлемое качество. Тем не менее, потеря либо полноты (при учете сокращений), либо точности (без их учета) составляет около 10%.

Статистический подход, основанный на контекстных векторах, позволяет существенно повысить качество. При использовании любого из трех протестированных модулей машинного обучения полнота и точность

083-Ury:_Layout 1 18.05.2008 16:25 Page 543

достигают 96-99%. Тест χ^2 показывает, что рост полноты статистически значим. Во втором эксперименте также статистически значимо и увеличение точности для всех программ машинного обучения. В первом эксперименте статистически значимый рост точности удалось достичь только с помощью SVM, но, по крайней мере, ни для C4.5, ни для Ripper не засвидетельствовано падения.

	Эксперимент 1		Эксперимент 2	
	точность, %	полнота, %	точность, %	полнота, %
termpunct	67.2	100**	66.9	98.9**
termpunct_cap	90.7	97.0**	89.6	96.0**
advanced	96.4	90.4	95.0	89.6
C4.5	97.8	98.5**	98.5*	97.5**
Ripper	98.5	98.5**	98.9**	96.0**
SVM-light	99.6**	98.5**	99.6**	97.5**

Таблица 3. Качество разбиения текста на предложения: результаты систем машинного обучения и контрольных эвристик. Показатели, значительно превосходящие полноту и точность эвристики `advanced`*, отмечены * (χ^2, $p<0.05$) и ** (χ^2, $p<0.01$)*

В таблице 4 приведены два примера работы нашего алгоритма (классификатор Ripper) в сравнении с эвристикой advanced. Как видно, статистический подход дает меньше ошибок. Особенно заметна разница при анализе узкоспециальных документов (нижняя половина таблицы). Слова «1» и «2» справедливо попали в список сокращений, в результате чего были пропущены две границы (с точки зрения эвристики advanced, вторая точка в «Рис. 1.» не отличима от первой). Этот пример демонстрирует главное преимущество статистического подхода к задаче сегментации на предложения по сравнению с эвристическими методами: для анализа узкоспециальных текстов достаточно просто добавить несколько соответствующих документов в обучающую выборку.

В то же время, полностью автоматический подход имеет и свом недостатки. Как показывает пример в верхней половине таблицы 4, наш автоматически составленный список сокращений (см. раздел 3) содержит достаточно много мусора. Например, слово «есть» было сочтено сокращением. Необходима лингвистическая экспертиза для ручной пост-обработки нашего списка.

Наконец, отметим, что и статистический алгоритм допускает ошибки. Наиболее проблематичными оказались сочетания точки и кавычек-лапок: несмотря на то, что часть таких контекстов проанализирована правильно, некоторые границы оказались пропущены, как в примере в верхней половине таблицы 4. В данный момент мы проводим дополнительные эксперименты по улучшению качества сегментации текстов с кавычками-лапками.

5. Заключение

В данной работе был предложен статистический подход к задаче определения границ предложений в произвольном тексте на русском языке. Наш алгоритм основан на контекстах знаков препинания и не требует синтаксического анализа, что позволяет обрабатывать документы с высокой скоростью.

Наши эксперименты показывают, что статистический подход позволяет добиться существенно более точного и полного выделения границ, чем наиболее распространенные эвристики.

Данная работа была проведена на материале газетных статей, отобранных случайным образом из Национального Корпуса Русского Языка. В будущем мы планируем изучить применимость нашего метода для обработки менее качественных текстов, прежде всего, веб-страниц. Наши предварительные исследования выявили целый ряд дополнительных задач, возникающих при анализе веб-документов.

Список литературы

1 http://ruscorpora.ru

2 Ровинская М. Точка как Проблема. Материалы Международной Конференции Диалог. 2000.

3 Quinlan J.R. C4.5: Programs for Machine Learning. Morgan Kaufman. 1993.

4 Cohen W.W. Fast Effective Rule Induction. Proceedings of ICML. 1995.

083-Ury:_Layout 1 18.05.2008 16:25 Page 544

5 Burges C.J. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2(2). 1998.

6 http://aot.ru

7 Reynar J.C. and Ratnaparkhi A. A Maximum Entropy Approach to Identifying Sentence Boundaries. Proceedings of ANLP, pp. 16-19. 1997.

8 Stevenson M. and Gaizauskas R. Experiments on Sentence Boundary Detection. Proceedings of ANLP-NAACL. 2000.

advanced	**статистический алгоритм (Ripper)**
Статья общего содержания (культура)	
<s> Был на церемонии момент , когда прозвучала пронзительно высокая и чистая нота .___" Ника " за " Честь и Достоинство "-- вот так , всё с заглавной буквы -- вручалась Петру Ефимовичу Тодоровскому .</s> *<s> Петру Тодоровскому -- оператору и режиссёру , композитору и музыканту , солдату и просто замечательному человеку .</s>* *<s> Он молодой , ошалевший от победной весны 45-го , смотрел на нас с экрана в хуциевском фильме " Был месяц май ".</s>* *<s> Он вышел на сцену под гром аплодисментов и " Рио-риту ".</s>* *<s> Для своих ровесников и друзей так и оставшийся в его - то годы Петей Тодоровским .</s>* *<s> Он прошёл через зал ," по главной улице с оркестром ", держа в руках гитару .</s>* *<s> Спасибо вам , дорогой Петр Ефимович !</s>* *<s> За веру , верность и " Верность ", за всё ваше кино , за то , что вы сделали для нас , за вашу нескончаемую любовь , за то , что вы есть .___За то , что " и всё-таки , и всё-таки , и всё-таки мы победили "!</s>* *<s> Той весной .</s>* *<s> За то , что у нас есть эта весна .</s>* *<s> И это ее семнадцатое мгновение .</s>*	*<s> Был на церемонии момент , когда прозвучала пронзительно высокая и чистая нота .___" Ника " за " Честь и Достоинство "-- вот так , всё с заглавной буквы -- вручалась Петру Ефимовичу Тодоровскому .</s>* *<s> Петру Тодоровскому -- оператору и режиссёру , композитору и музыканту , солдату и просто замечательному человеку .</s>* *<s> Он молодой , ошалевший от победной весны 45-го , смотрел на нас с экрана в хуциевском фильме " Был месяц май ".</s>* *<s> Он вышел на сцену под гром аплодисментов и " Рио-риту ".</s>* *<s> Для своих ровесников и друзей так и оставшийся в его - то годы Петей Тодоровским .</s>* *<s> Он прошёл через зал ," по главной улице с оркестром ", держа в руках гитару .</s>* *<s> Спасибо вам , дорогой Петр Ефимович !</s>* *<s> За веру , верность и " Верность ", за всё ваше кино , за то , что вы сделали для нас , за вашу нескончаемую любовь , за то , что вы есть .</s>* *<s> За то , что " и всё-таки , и всё-таки , и всё-таки мы победили "!</s>* *<s> Той весной .</s>* *<s> За то , что у нас есть эта весна .</s>* *<s> И это ее семнадцатое мгновение .</s>*
Статья технического содержания	
<s> Чтобы объективно оценивать качество изоляции , используя явление абсорбции , в Нижегородском филиале РГОТУПСа было разработано оригинальное устройство , принципиальная схема которого приведена на рис . 1 .___Оно включает в себя : высоковольтный стабилизированный источник питания ВИП с выходным напряжением 1000 или 2500 В , измерители тока И 1 и напряжения И 2 , два высоковольтных реле Р 1 и Р 2 .___Последними управляет микроЭВМ или система автоматики , построенная на интегральных микросхемах с применением программируемых запоминающих устройств .</s> *<s> Схема замещения неоднородной изоляции тягового двигателя представлена в виде двух конденсаторов С 1 и С 2 , зашунтированных резисторами R1 и R2.</s>*	*<s> Чтобы объективно оценивать качество изоляции , используя явление абсорбции , в Нижегородском филиале РГОТУПСа было разработано оригинальное устройство , принципиальная схема которого приведена на рис . 1 .</s>* *<s> Оно включает в себя : высоковольтный стабилизированный источник питания ВИП с выходным напряжением 1000 или 2500 В , измерители тока И 1 и напряжения И 2 , два высоковольтных реле Р 1 и Р 2 .</s>* *<s> Последними управляет микроЭВМ или система автоматики , построенная на интегральных микросхемах с применением программируемых запоминающих устройств .</s>* *<s> Схема замещения неоднородной изоляции тягового двигателя представлена в виде двух конденсаторов С 1 и С 2 , зашунтированных резисторами R1 и R2.</s>*

Таблица 4. Сегментация на предложения с помощью эвристики advanced (левая колонка) и статистического алгоритма (правая колонка). Границы предложений обозначены тэгом <s>..</s>. Ошибки подчеркнуты и выделены серым.

084-Fomchenko_Azarova:_Layout 1 18.05.2008 16:38 Page 545

ВЗАИМОДЕЙСТВИЕ ЭСТЕТИЧЕСКИХ, МОРАЛЬНЫХ И ПРАГМАТИЧЕСКИХ АСПЕКТОВ В СЕМАНТИЧЕСКОЙ СТРУКТУРЕ ОЦЕНОЧНЫХ ПРИЛАГАТЕЛЬНЫХ РУССКОГО ЯЗЫКА

INTERACTION OF PRAGMATIC, AESTHETIC AND MORAL FEATURES IN THE SEMANTIC STRUCTURE OF RUSSIAN JUDGMENT ADJECTIVES

Фомченко А.В. *(degteva.anna@gmail.com),* ***Азарова И.В.*** *(ivazarova@gmail.com)*
Санкт-Петербургский государственный университет

В докладе рассматривается ядро группы оценочных прилагательных в русском языке и особенности их представления в компьютерном тезаурусе RussNet. Обсуждаются базовые аспекты оценочного значения – прагматический, эстетический и моральный – и их взаимодействие для частотных и низкочастотных прилагательных.

1. Введение

На Кафедре математической лингвистики Санкт-Петербургского государственного университета создается компьютерный тезаурус RussNet для современного русского языка [2], который следует основным принципам, разработанным при построении Принстонского WordNet [16] и других wordnet-словарей.

Wordnet-словари предоставляют широкие возможности для отображения семантических отношений между значениями, получившими лексикализованное выражение в рамках некоторого языка. Среди отношений есть традиционные для тезаурусной организации лексических значений, например, родовидовые, или гипонимические, а также менее традиционные: меронимические (часть-целое), каузативные, различные семантико-деривационные отношения [7]. Однако в последнее время стало ясно, что необходима особая схема для отображения различных аспектов оценочного значения, в первую очередь, для автоматического выявления оценочного компонента в текстах [14] (opinion mining, sentiment mining), или так называемой тональности текста [6].

Проблема автоматического определения субъективного или оценочного компонента для слов, предложений и текстов в последнее время получила достаточное отражение в научной и практической литературе, можно упомянуть статьи [8, 9, 10, 11, 12], представленные на специализированном семинаре в Сиднее в 2006 г. Они показывают, что (1) субъективность текстов и представленные в них оценки довольно сложно отобразить простым способом; (2) первоначальный подход к оценочному компоненту как простой «поляризации» положительных и отрицательных текстов являлся необходимым начальным этапом упрощения задачи для выработки более сложной методики определения тональности текстов; (3) в настоящее время по-прежнему не совсем ясно, как выглядит шкала оценок, которые могут быть выражены в текстах.

2. Структура оценочных значений

Говоря о субъективном или оценочном компоненте в тексте, необходимо подчеркнуть, что помимо собственно оценки (положительной, отрицательной или отсутствия оценки – нейтральности), имеется объект оценки (что оценивается) и субъект оценки (кто оценивает). Как это ни парадоксально, второй и третий компоненты регулярно упускают из виду, поскольку наибольшее количество работ по автоматическому определению оценки используют данные сайтов, на которых покупатели или пользователи оценивают ту или иную продукцию. В таких текстах объекты оценки более-менее однородны, а оценка анонимна, но подчеркивается [10], что даже в таких текстах могут высказываться определенные соображения (*Я считаю, что этот товар должен стоить около 15$*), которые могут интерпретироваться как оценочные. Вообще, некоторое логическое основание оценочного компонента, когда можно привести доводы в пользу или против чего-либо, на базе которых и производится оценка объекта/ лица субъектом [8], является наиболее адекватным представлением «тональности» в автоматических системах.

Несмотря на то, что довольно большое количество практических работ по определению тональности выполнены на базе систем машинного обучения, в настоящее время становится очевидным, что понимание того, на основании чего принимается то или иное решение в системе анализа текста, является более важным и амбициозным направлением работ. В любом случае, практически все описания систем определения

084-Fomchenko_Azarova:_Layout 1 18.05.2008 16:38 Page 546

тональности текстов приводят примерно одни и те же этапы выполнения работ, среди которых центральное место занимает определение слов, передающих субъективно-модальное или оценочное значение. Обычно в группу наиболее важных для данного аспекта значений включают глаголы и прилагательные [11] или прилагательные и существительные [6]. Причем в последнем случае, очевидно, есть определенный приоритет (или доминирование) оценочного статуса прилагательных над существительными, в частности, положительная оценка в значении существительного (*защитник демократии*) подавляется отрицательной оценкой прилагательного (*плохой/ отвратительный защитник демократии*), но не наоборот (например, **замечательный диктатор*, **отличный жулик*).

Вообще, существует представление о большем весе или разнообразии отрицательных характеристик, которое вроде бы подкрепляется тем, что отрицательные значения чаще находят выражение в текстах [15], имеют большое лексическое разнообразие, в пример довольно часто приводят даже ироничное инвертирование положительной оценки лексем повышенного статуса в разговорной речи: *не соизволите ли убрать за собой грязь*.

Исследование оценочного значения существительных, обозначающих лица, которое было выполнено на кафедре математической лингвистики в русле заявленной темы [5], показало, что тексты, описывая одну и ту же ситуацию (специально была выбрана ситуация взятия заложников, поскольку она вряд ли предполагает безучастную, «нетональную» передачу), могут весьма значительно варьировать степень тональности. Нейтральные слова (например, *мужчина*) могут довольно неожиданно получить отрицательную оценку в контексте ситуативного действия (когда мужчины выбегали из помещения школы, оставив детей).

В данной работе нами будут исследованы значения прилагательных, которые безусловно участвуют в формировании субъективной окраски текстов. В первую очередь, мы акцентируем собственно структуру значений «оценочных» прилагательных, на основании которых можно предложить шкалу или пространство оценочных значений.

Аспекты классификации оценочных значений включают: модальные и дескриптивные, абсолютные и сравнительные, зависимые и независимые от контекста [4]. В центре нашего внимания будут эстетические, моральные и прагматические значения прилагательных, реализующие компонент частнооценочных значений, которые создают особую таксономию объектов [3], отличную от природной. Целью исследования является выявление объема пересечения этих аспектов оценки и их взаимодействие в семантической группе оценочных прилагательных в русском языке.

3. ЛСГ оценочных прилагательных в русском языке

Для выявления структуры лексико-семантической группы оценочных прилагательных рассмотрим значения наиболее частотных представителей этой группы: *хороший, красивый, правильный, удобный, приятный, прекрасный, плохой и человеческий,* кроме того, будут исследованы менее частотные *милый, уродливый и некрасивый, неудобный, неправильный, неприятный, дурной,* чтобы изучить особенности в распределении положительных и отрицательных компонентов оценки.

Значения этих слов, представленные в Словаре русского языка в четырех томах под редакцией А.П. Евгеньевой (М., 1981) − далее МАС, используются для первоначальной разметки значений контекстов в корпусе современных текстов кафедры математической лингвистики объемом в 21 млн. словоупотреблений [1]. Это стандартная методика определения структуры значений для некоторой лексической группы в RussNet [1], которая позволяет экстраполировать результаты разметки случайной выборочной совокупности контекстов (100−150) на корпус в целом, выявляя таким образом реальную структуру распределения значений в лексико-семантической группе.

В результате проверки значений перечисленных слов по корпусу было выделено 100 оценочных значений с общим объемом контекстов 850,31 ipm[1].

3.1. Ядро и периферия группы оценочных прилагательных

Центральной лексемой группы является прилагательное *хороший*$_1$[2], которое выражает, в первую очередь, прагматическую оценку, что подчеркивается значением «…вполне отвечающий своему назначению», с частотностью 103 ipm: *хорошее вино, хороший вкус, хороший город, хороший друг, хороший доход, хороший закон, хорошая зарплата, хорошее здоровье, хорошее зрение, хорошее качество* и проч. При всем многообразии объектов, получающих такую характеристику, основной чертой является пригодность, соответствие критериям

[1] ipm – instanses per million, имеется в виду число употреблений на миллион словоупотреблений в корпусе.
[2] Номера значений прилагательных в статье соответствуют их частотному распределению в корпусе.

084-Fomchenko_Azarova:_Layout 1 18.05.2008 16:38 Page 547

качества, отсутствие отрицательных черт, при этом ослабляется собственно положительная оценка, акцентируется соответствие стандарту: *хорош для детского питания, если морковь уже хороша*. Это значение в сочетании с интенсификацией положительных свойств («идеально подходящий для чего-либо») и некоторым сужением спектра значений реализовано у прилагательного *прекрасный$_3$* (8 ipm) (*прекрасное лекарство*; *дает прекрасную протраву*); это значение отлично от значения *прекрасный$_1$* (43 ipm) «очень хороший; превосходный», имеющего дополнительную экспрессивную оценку. Высокая частотность *хороший$_1$* в корпусе (> 100 ipm) указывает на его центральный характер в иерархии оценочных признаков (в концепции wordnet-словарей такие значения относятся к basic concepts – основным понятиям). Примыкают к ядерному значению на этой шкале прилагательные, по-разному акцентирующие прагматическое значение («такой, какой нужен», «такой, которым удобно и легко пользоваться» и проч.): *приятный$_3$* (35,0 ipm) (*приятное место, приятное время*), *правильный$_1$* (17,1 ipm), *удобный$_1$* (16,9 ipm). Нужно отметить, что сочетания существительных с *хороший$_1$* частично синонимичны сочетаниям с прилагательными, примыкающими к ядру: *хорошее = приятное место, хороший = правильный закон, хорошее = удобное устройство*, для некоторых сочетаний синонимичными будут связанные выражения, «лексические функции»: *хороший = верный друг, хорошая = высокая зарплата, хорошее = крепкое здоровье*. Группа прагматических оценочных значений является самой многочисленной (см. Таблицу 1) и занимает самую большую долю (55%) среди рассмотренных контекстов корпуса (см. Таблицу 2).

Доля контекстов, выражающих морально-нравственную оценку, составляет 30,4%, то есть стоит на «втором месте» в рассматриваемой группе, однако базовое прилагательное *красивый* имеет более очевидный статус в группе эстетических оценок благодаря суммарной частотности значений *красивый$_1$* и *красивый$_2$* (71,2 ipm), причем второе значение отличается от первого типом объекта, подвергающегося эстетической оценке – человека, точнее человеческого лица. Если обратиться к структуре значений базового оценочного слова *хороший*, то морально-нравственное значение также представлено в ней как более значимое: *хороший$_3$* (19,2 ipm) «обладающий положительными моральными качествами» в отличие от *хороший$_9$* (7,2 ipm) «очень красивый» (*хороша собой*). Морально-нравственное значение *хороший* реализуется в сочетании с объектами, обозначающими лица и группу лиц, – это морально-нравственная оценка поведения человека в семье и социуме: *хороший человек, хорошая мать, хороший муж,* примыкает к этой группе оценка профессиональных качеств: *хороший учитель, хороший консультант, хороший музыкант, хороший попутчик, хороший коллектив*.

Значение	**1**	**2**	**3**
хороший1	(+)103,0		
прекрасный1	(+)43,0		
приятный3	(+)35,0		
человеческий3	(+)23,0		
правильный1	(+)17,1		
хороший4	(+)17,0		
удобный1	(+)16,9		
правильный2	(+)14,3		
удобный2	(+)12,5		
правильный3	(+)10,1	(+)10,1	
хороший5	(+)10,0		(+)10,0
удобный3	(+)9,9		
хороший6	(+)9,6		
удобный4	(+)8,8		
прекрасный3	(+)8,0		
хороший7	(+)7,2		
добрый7	(+)4,5		
добрый6	(+)4,5	(+)4,5	
правильный4	(+)4,2		
добрый8	(+)3,9		
приятный1	(+)2,5		
правильный6	(+)2,4		
хороший13	(+)2,0		
добрый11	(+)1,3		
правильный9	(+)0,7		
правильный8	(+)0,7		

Значение	**1**	**2**	**3**
плохой1	(–)31,5		
неприятный2	(–)8,9		
неприятный3	(–)7,3		
неправильный1	(–)7,3		
плохой5	(–)5,0		
плохой4	(–)5,0		
хороший11	(–)4,8		
плохой7	(–)4,3		
дурной4	(–)4,1		
дурной3	(–)4,1		
неправильный2	(–)3,8		
неправильный3	(–)2,8		
уродливый1	(–)2,7	(–)2,7	
плохой8	(–)2,2		
неудобный1	(–)1,9		
неудобный3	(–)1,1		
неудобный2	(–)1,1		
дурной6	(–)1,1		
плохой9	(–)0,7		
неудобный4	(–)0,4		
добрый1		(+)37,9	
человеческий2		(+)24,5	
хороший2		(+)21,6	
хороший3		(+)19,2	
приятный4		(+)16,0	(+)16,0
добрый2		(+)14,1	

084-Fomchenko_Azarova:_Layout 1 18.05.2008 16:38 Page 548

Значение	**1**	**2**	**3**
добрый4		(+)9,6	
милый1		(+)6,0	
добрый5		(+)5,7	
милый2		(+)4,9	
хороший10		(+)4,8	
милый3		(+)3,7	
добрый9		(+)3,2	
красивый4		(+)3,0	
добрый10		(+)2,6	
хороший12		(+)2,4	
правильный7		(+)2,1	
хороший14		(+)2,0	
милый4		(+)2,0	
милый6		(+)0,9	
добрый12		(+)0,6	
милый7		(+)0,1	
неприятный1		(–)13,7	
плохой2		(–)10,7	
плохой3		(–)6,5	
дурной1		(–)4,8	(–)4,8
дурной2		(–)4,6	
плохой6		(–)4,3	
неприятный4		(–)3,2	
дурной5		(–)3,0	

Значение	**1**	**2**	**3**
неправильный5		(–)1,0	
дурной7		(–)0,7	
красивый5		(–)0,6	
неправильный6		(–)0,5	
некрасивый3		(–)0,4	
неудобный5		(–)0,2	
красивый1			(+)41,9
красивый2			(+)29,3
прекрасный2			(+)26,0
хороший9			(+)7,2
прекрасный4			(+)5,0
красивый3			(+)3,5
правильный5			(+)3,3
милый5			(+)0,9
некрасивый1			(–)2,6
уродливый2			(–)2,1
некрасивый2			(–)1,4
неправильный4			(–)1,3
уродливый3			(–)1,0
некрасивый4			(–)0,1
добрый3	10,3		
хороший8	7,2		
правильный10	0,4		
приятный2	0,2		

Таблица 1. Распределение оценочных значений по группам в ipm
(1 – прагматическое, 2 – морально-нравственное, 3 – эстетическое значения;
(+) отмечает положительное, а (–) – отрицательное значения).

Центральным в группе морально-нравственных оценок является значение *добрый*$_1$ (37,9 ipm), к которому ближе всего *человеческий*$_2$ (24,5 ipm) «такой, какой должен быть принят у людей, какой подобает людям»: *человеческое состояние, человеческие условия* и т. д. Это прилагательное может передавать отрицательную моральную оценку (*трусливая и жалкая природа человеческая*), что, очевидно, связано с оценочным значениям существительного: *человеческие слабости* vs *человеческие тепло, человеческие ценности*, когда интерпретацию получает словосочетание целиком, при этом отметим, что значение *человеческий*$_1$ в основном нейтрально: *человеческое общество, человеческий язык*. К ядерной группе положительных атрибутивных морально-нравственных оценок примыкает *приятный*$_4$ (16 ipm), *правильный*$_3$ (10 ipm), первое значение совмещено с эстетическим, второе − с прагматическим значением.

Значение прилагательного *милый* представляет особый интерес, так как его значение находится на стыке двух оценок: содержащаяся в нем оценка внешности не является чисто эстетической (не говорит о красоте черт) или чисто нравственной, значение может быть определено как «обладающий внешним видом, вызывающим симпатию, говорящем о положительных душевных качествах». Следует отметить, что основное различие между значениями *милый*$_2$ «располагающий к себе» (*милое лицо*) и *милый*$_1$ «близкий сердцу; родной, дорогой» (*милые мне края*) состоит в принадлежности оценки. В первом случае это собственное свойство объекта оценки; во втором случае − описывается отношение к объекту другого лица. Данное прилагательное чаще всего служит для передачи отношений привязанности и симпатии, и в меньшей степени для описания эстетических свойств. Второе значение более частотно, и зачастую используется в качестве сравнительной оценки, приобретая значение «лучше».

Прилагательное *красивый*$_4$ со значением «отличающийся полнотой и глубиной внутреннего содержания» передает морально-этическую оценку лица или его действий. Пятое значение данного слова «рассчитанный на эффект, на внешнее впечатление», напротив, содержит отрицательную морально-нравственную оценку, направленную однако не столько на сам объект оценки (*красивые слова*), сколько на лицо, совершающее некоторое действие или поступок.

Группа отрицательных значений представлена, в первую очередь, значениями прилагательного *плохой*, первое значение (31.5 ipm) является антонимом *хороший*$_1$ и содержит в себе прагматический компонент

084-Fomchenko_Azarova:_Layout 1 18.05.2008 16:38 Page 549

«негодности». Большая область пересечения морально-нравственных и эстетических характеристик наблюдается для слова *уродливый*. Первое словарное значение («имеющий физические недостатки») является самым низкочастотным (1 ipm), а два основных – прагматическое, совмещенное с моральной оценкой, и чисто эстетическое «очень некрасивый, безобразный» (*уродливые строения*). Большая часть значений *неправильный* характеризует несоответствие норме (неправильный визуально, несимметричный) и в связи с этим имеет оттенок отрицательной оценки, направленной на эстетические свойства объекта, хотя доминирует прагматически обусловленное значение «не такой, какой нужен» (*неправильный столик, неправильное слово*).

В приведенной ниже таблице указаны доли контекстов с прагматической, морально-нравственной и эстетической оценками для рассмотренных прилагательных.

	1	2	3	Итого	%
Положительные	372,1	201,5	143,1	676,1	79,5
Отрицательные	100,1	56,9	13,3	120,9	14,2
Итого	472,2	258,4	156,4	850,3	100,0
%	55,5	30,4	18,4		

Таблица 2. Распределение долей оценочных значений по группам в ipm и процентах (1 – прагматическое, 2 – морально-нравственное, 3 – эстетическое значения).

4. Выводы и перспективы исследования

Опираясь на исследованную группу прилагательных с прагматическими, эстетическими и морально-нравственными значениями, можно отметить, что первая группа − прагматические значения − является самой представительной. В ядре этой группы находятся базовые значения самого частотного прилагательного *хороший*, которое выражает также и другие виды оценки. Относительно самостоятельный характер имеет группа значений, передающих эстетическую оценку. В этой группе центральное положение занимает прилагательное *красивый*. Группа морально-нравственных значений имеет наибольшее количество пересечений с другими группами, что, вероятнее всего, свидетельствует о ее развивающемся характере.

Интересно, что число контекстов, передающих отрицательное оценочное значение, составляет лишь незначительную часть (14,2%) от общего числа таких употреблений. Доля нейтральных контекстов в рассмотренной совокупности контекстов составляет примерно половину (52%) от общего количества.

Можно сформулировать для дальнейшей проверки гипотезу о том, что прагматическая оценка (в совокупности своих аспектов) и эстетическая оценка составляют базу для других оценочных значений, в частности, морально-нравственных, именно в этой сфере происходит параллельное развитие полярных оценочных значений.

Список литературы

1. Азарова И.В., Синопальникова А.А. Использование статистико-комбинаторных свойств корпуса современных текстов для формирования структуры компьютерного тезауруса RussNet // Труды международной конференции «Корпусная лингвистика 2004». 11−14 октября 2004 г. СПб., 2004. С. 5−15.

2. Азарова И.В., Синопальникова А.А., Яворская М.В. Принципы построения wordnet-тезауруса RussNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог’2004 («Верхневолжский», 2−7 июня 2004 г.) М., 2004. С. 542−547.

3. Арутюнова Н.Д. Аксиология в механизмах жизни и языка // Проблемы структурной лингвистики, 1982. М., 1984. С. 5−23.

4. Вольф Е.М. Функциональная семантика оценки. М., 2002.

5. Дегтева А.В. Особенности функционирования наименований лиц в текстах, освещающих конфликтные ситуации (на материале статей, посвященных захвату заложников) // Материалы XXXVI международной филологической конференции. Вып. 10. Прикладная и математическая лингвистика. СПб., 2007. С. 28−34.

6. Ермаков А.Е., Киселев С.Л. Лингвистическая модель для компьютерного анализа тональности публикаций СМИ // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог’2005. М., 2005.

084-Fomchenko_Azarova:_Layout 1 18.05.2008 16:38 Page 550

7. Azarova I., Mitrofanova O., Sinopalnikova A., Yavorskaya M., Oparin I. RussNet: Building a Lexical Database for the Russian Language // Workshop on WordNet Structures and Standardisation, and how these affect Wordnet Application and Evaluation. 28th May 2002. Las Palmas de Gran Canaria, 2002. P. 60–64.

8. Fujii A., Ishikawa T. System for Summarizing and Visualizing Arguments in Subjective Documents: Toward Supporting Decision Making // Proceedings of the ACL/COLING Workshop on Sentiment and Subjectivity in Text. Sydney, Australia. P. 15−22.

9. Généreux M., Evans R. Towards a validated model for affective classification of texts // Proceedings of the ACL/COLING Workshop on Sentiment and Subjectivity in Text. Sydney, Australia. P. 55−62.

10. Hiroshima N., Yamada S., Furuse O., Kataoka R. Searching for Sentences Expressing Opinions by using Declaratively Subjective Clues // Proceedings of the ACL/COLING Workshop on Sentiment and Subjectivity in Text. Sydney, Australia. P. 39−46.

11. Kim S.-M., Hovy E. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text // Proceedings of the ACL/COLING Workshop on Sentiment and Subjectivity in Text. Sydney, Australia. P. 1−8.

12. Stoyanov V., Cardie C. Toward Opinion Summarization: Linking the Sources// Proceedings of the ACL/COLING Workshop on Sentiment and Subjectivity in Text. Sydney, Australia. P. 9−14.

13. The Global WordNet Association // URL: http://www.globalwordnet.org/

14. Tufiş, D., Ion, R.: Cross lingual and cross cultural textual encoding of opinions and sentiments. Semantics, Opinion and Sentiment in Text. Iaşi, July 23–August 3, 2007.

15. Wierzbicka A. Ethno-syntax and the Philosophy of Grammar // Studies in Language. 1979. № 3.

16. WordNet: An Electronic Lexical Database / Ch. Fellbaum (ed.). MIT Press, 1998.

085-Cimmerling_A:_Layout 1 18.05.2008 21:09 Page 551

ЛОКАЛЬНЫЕ И ГЛОБАЛЬНЫЕ ПРАВИЛА В СИНТАКСИСЕ
LOCAL AND GLOBAL RULES IN SYNTAX[1]

Циммерлинг А.В. *(meinmat@yahoo.com)*
МГГУ им. М.А.Шолохова, Российский государственный гуманитарный университет

Предметом внимания настоящей статьи являются феномены порядка слов и фразовой просодии, доступные для изучения на разных уровнях и описываемые при помощи правил двух типов – локальных и глобальных. Сочетание данных типов правил характерно для многоуровневых моделей языка и для алгоритма порождения структурных объектов в формальных грамматиках. Действующий в русском языке механизм выбора носителя фразового акцента зеркально отражает иерархию членов предложения и может быть задан в виде локального правила. Порядок слов в русском языке регулируется Линейно-Акцентными преобразованиями (ЛА-преобразований), одновременно меняющими место и акцентную маркировку коммуникативных составляющих. Предлагается список из 7-8 ЛА-преобразований, позволяющих строить исчисление русских предложений с уже приписанными темами и ремами, и обсуждается возможности задать правила ЛА-преобразований способом, совместимым с Мягко Контекстно-Свободными Грамматиками.

При формализации естественных языков и при моделировании формальных грамматик часто возникает необходимость применения правил двух типов. Правила первого типа – назовем их локальными – порождают замкнутые языковые объекты, которые остаются непроницаемы или, по крайней мере, сохраняют присущие им категориальные свойства (categorical features) при включении их на дальнейшей стадии порождении языка в объекты большей длины или более высокого иерархического уровня. Правила второго типа – назовем их глобальными – применяются позже: они порождают сложные объекты, составные части которых уже заданы локальными правилами, одновременно регулируя те нетривиальные свойства элементарных объектов, которые зависят от конфигурации сложных объектов и не могут проявиться на начальной стадии порождения. Неформально это означает, что, например, правила построения словоформ языка L можно сформулировать так, что они не будут зависеть от правил построения словосочетаний в этом языке, но правила построения словосочетаний в языке L могут обращаться к нетривиальным свойствам словоформ, не являющихся общими для всех словоформ языка L. Сходным образом, правила линейного расположения главного (head) и зависимого (complement) компонента словосочетаний в языке L можно задать так, что они не будут зависеть от структуры предложения (clausal structure)[2], но правила построения предложения в этом языке будут влиять на перестановку или разрыв этих компонентов[3]. Этот ряд можно продолжить, если представить порождение линейно упорядоченного дерева предложения в языке L и присваивание элементам предложения коммуникативных статусов темы и ремы как последовательные стадии порождения[4]. Бесконечно долго данный прием эксплуатировать нельзя по двум причинам. Во-первых, набор единиц и уровней в естественном языке ограничен. Во-вторых, построение особого модуля для каждого новой вводимой в рассмотрение группы факторов подрывает вычислительную эффективность (computational efficiency) лингвистического процессора, превращая его из работающего устройства в демонстрацию той или иной программы описания языка, ср. сходные соображения в (Gärtner & Michaelis 2007, 177).

Нет уверенности в том, что вся лингвистическая информация, полученная при порождении структурного объекта на выходе (output) n-уровня, будет релевантна на входе (input) следующего, n+1 уровня. Многие авторитетные ученые (Н.Хомский, И.А.Мельчук и т.п.) готовы дать положительный или отрицательный ответ на этот вопрос, ср. (Chomsky 2005, 11) но их мнение мотивировано либо избранным инженерным решением – как

[1] Статья написана при финансовой поддержке гранта РГНФ 06-04-00203а “Типология языков со свободным порядком слов и модели инверсии”. Автор выражает благодарность А.А.Перекрестенко и Т.Е.Янко за ценные замечания.

[2] Это возможно при условии, что модуль языка L, ответственный за порождение предложений, получает на входе (input) уже упорядоченные словосочетания и не может работать с линейно неупорядоченными цепочками слов.

[3] Это возможно при условии, что грамматика языка L разрешает реструктуризацию уже порожденных и линейно упорядоченных структур. Адекватное описание порядка слов в естественных языках без этого условия, по-видимому, исключено.

[4] Ср. пионерскую работу (Падучева 1984), где намечена идеология такого подхода.

085-Cimmerling_A:_Layout 1 18.05.2008 21:09 Page 552

формализовать естественный язык, либо отстаиваемой лингвистической доктриной (если она не сводится целиком к последней задаче). В этих условиях обсуждение реально действующих механизмов, поддающихся описанию в терминах комбинации локальных и глобальных правил построения, может пролить свет на то, являются ли эти правила фантомом, поддерживаемым алгоритмическим уровневым подходом к языку, или же за ними стоит языковая реальность. Перед тем как перейти к непосредственному предмету нашей статьи – правилам выбора носителя главного фразового акцента и линейно-акцентным преобразованиям в современном русском языке – кратко рассмотрим употребления терминов «локальный» vs «глобальный» в лингвистической литературе. Подборка не претендует на полноту, а очередность цитат не соотнесена с частотой тех или иных употреблений термина.

Д.Чавар и К. Уайльдер (Ćavar, Wilder 1999, 462-463), описывающие правила расстановки клитик и глагола в (сербо)-хорватском языке, квалифицируют часть этих правил как «локальные» в плане Минималистской Программы Хомского, так как они *применяются сразу, как только возникают условия /мишени* (targets) для их применения[5]. В данном употреблении «локальное правило» значит то же, что «циклическое» или «рекурсивное» правило[6]. Хотя выводы Чавара-Уайлдера вызывает сомнения, локальные$_1$ = циклические правила линеаризации предложения в языках мира найти можно. В синтаксисе словенского языка правило постановки сентенциальных клитик на второе место действует циклически, снизу вверх по дереву предложения (down-top). Словенские клитики, в отличие от сербохорватских, могут присоединяться к сентенциальным составляющим, в составе которых уже могут быть другие клитики, упорядоченные при предыдущем применении того же правила. В примере (1) цепочка энклитик главного предложения *se (1)= je = (2)= že = (3)* примыкает справа к вынесенному в начало конструкции придаточному, в составе которого есть другая цепочка клитик *=smo* (1) *=se* (2), примыкающая справа к подчинительному союзу ko “когда”. Еще более показателен пример (2), где клитика je в составе главного предложения примыкает справа к развернутой обстоятельственной группе вида [AdvP1 [CP 2]], куда вставлено придаточное времени [$_{CP}$ kjer *=je* komponiral], внутри которой есть своя клитика. Тем самым, клитики главных предложений в примерах (1), (2), в точном соответствии с данным выше определением «локального$_1$ правила», упорядочиваются, как только достраивается очередной фрагмент синтаксической структуры, содержащий подходящую «мишень» для клитик, т.е. полную начальную составляющую, к которой клитика может присоединяться[7].

(1) слов. #[[ko = *smo* (1) *=se* (2) vrnili] = *se* (1) = *je* (2) = *že* (3) stemnilo]
букв. #[[когда *=мы* (1)= *ся* (2) вернули] = *ся* =(1) *есть* (2)= *уже* (3) смеркало]

(2) Слов.

XP = [[AdvP1 [CP 2]]	Cl	Sub	VP
[Po izgubi svojega domicila v Steinbachi am Attersee [kjer (1) = je (2) komponiral]]	**Je**	Gustav Mahler	Iskal miren kraj
[После утраты своего пристанища в Стейнбахе на (озере) Аттерзее [где (1) =связка-3 л.ед.ч. (2) он-сочинял музыку]]	**Связка-3 л. ед.ч.**	Густав Малер	Искал тихое место

Я.Г.Тестелец в статье (Testelets 2003) критикуя гипотезу А.Кардиналетти и М.Штарке (Cardinaletti, Starke 1999) о предположительно универсальном различении т.н. сильных (постоянно ударных) и т.н. слабых (атонируемых) местоимений, замечает, что свойства, которые Кардиналетти и Штарке считают для слабых местоимений диагностическими, например, неспособность неодушевленных местоимений сочиняться в контекстах типа рус. <*путеводитель$_i$ и брошюра$_j$ на столе*>.* *Возьми **его**$_i$ и **ее**$_j$*., объясняются «нелокальным фактором», т.е. особенностями этих контекстов/ общим прагматическим принципом, а не собственными структурными свойствами [гипотетических] слабых местоимений (Testelets 2003, 521). В данном отрывке «локальный» значит «присущий р по определению», «ингерентный», а антитеза к локальный$_2$ = ингерентный – передает значения «привходящий», «привносимый контекстом, в который включено р»: из комментариев Я.Г.Тестельца следует, что «нелокальными$_2$» будут правила, требующие обращения к несобственно синтаксическим характеристикам – подбору лексики и информационной структуре высказывания[8].

[5] ‘Pre-Spell-Out operations are local, in the sense that an operation must apply as soon as its target is created’ – (Ćavar, Wilder 1999, 462).

[6] Термин «циклический», и в самом деле, присутствует на соседней странице статьи (Ćavar, Wilder 1999, 463).

[7] Как ни странно, сторонниками Хомского данная особенность словенского языка пока не освоена. С точки зрения хомскианца “мишенью” (target) для словенских клитик является не наличие внешне выраженной (spelled-out) начальной группы, а наличие пустой позиции C^0, которая притягивает клитики. Поскольку, однако, эта позиция возникает тогда, когда заполняется начальная позиция (SpecCP/SpecTP), формулировка, данная выше, и ее реинтерпретация в хомскианских терминах экстенсионально эквивалентны.

085-Cimmerling_A:_Layout 1 18.05.2008 21:09 Page 553

Иначе использует термин «локальный» И.А.Мельчук в своей классификации поверхностно-синтаксических отношений в бинарных группах вершина (head) + зависимое (complement). Он выделяет три случая: (1) и положение лексической вершины, и положение зависимого слова фиксированы. Ср. группы «предлог → существительное» в русском языке: *на стол / на красивый стол / на стол брата*. (2) Задано направление развертывания зависимого слова, но не его место в группе. Ср. рус. *Автобус на Петербург фирмы «Финнорд» ~ Автобус фирмы «Финнорд» на Петербург.* (3) Направление развертывания не задано. Ср. группы «сказуемое → подлежащее» и «сказуемое → обстоятельство» в русском языке. Правила типа (1) локальные и бесконтекстные. Правила типа (2) – их И.А.Мельчук называет «квазилокальными» – требуют для упорядочения A → B рассмотрения всех других групп, которые могут зависеть от A, например A → Z, A → W, и определения порядка B.... Z, B.... W. Правила типа (3) по И.А.Мельчуку, «глобальные«, поскольку они требуют для упорядочения A → B рассмотрения всей фразы (Мельчук 1995: 278). В данной классификации термин «локальный» является синонимом «контекстно-свободный», а антитезой локальный$_3$=контекстно-свободный будут разные типы контекстно-зависимых правил. По-видимому, «квазилокальные» правила И.А.Мельчука, предполагающие возможность вставки элементов между A и → B, можно интерпретировать как Мягко Контекстно-Зависимые (Mildly Context-Sensitive) в смысле Джоши (Joshi 1985), а его «глобальные» правила являются сильно Контекстно-Зависимыми (Context-Sensitive). Для самого И.А.Мельчука «глобальность» означает необходимость просматривать всю синтаксическую область, в то время как «квазилокальность» относится к случаям, где для упорядочения фрагмента X достаточно ограничиться просмотром другого фрагмента Y. По нашему мнению, употребления локальный$_3$= КС-свойство и локальный$_2$=ингерентное свойство близки: если задать один из встречающихся вариантов (в данном примере, разновидность порядка слов) в качестве основного (default option), то правила типа (2) и (3) удастся свести друг к другу[9].

Четвертое употребление термина «локальный» утвердилось в теории формальных языков (Stabler 1997; Stabler 1998) под влиянием ранних работ Д.Росса (Ross 1967) и Н.Хомского (Chomsky 1977). Оно передает идею непроницаемости рабочего участка синтаксической структуры для операций, которые могут иметь место в пределах более широкого участка, его перекрывающего. Разные комбинации локальных$_4$ ограничений (Locality Conditions) дают разные классы формальных грамматик, что показано в (Gärtner & Michaelis 2007, 171-176). Пересечения локальный$_4$ = непроницаемый для последующих операций с рассмотренными ранее употреблениями термина, особенно с локальный$_1$и локальный$_3$, очевидны, тем более, что в списке локальных$_4$ ограничений есть запреты на дальнее перемещение элемента y в позицию U при наличии ближестоящего элемента x, способного переместиться в U (Short Move Condition), вынос элемента из группы, занимающей позицию подлежащего (Specifier Island Constraint) и вынос элемента из группы, занимающей позицию адъюнкта (Adjunct Island Constraint), т.е. на явления, нарушающие локальность$_1$ и локальность$_3$.

Значение	Интерпретация	Типичный контекст применения
Локальный$_1$	Циклический	Алгоритмизация разных правил/актов применения одного правила
Локальный$_2$	Ингерентный	Добавление новых факторов, влияющих на уже построенную структуру
Локальный$_3$	Контекстно-свободный	Сужение/расширение синтаксической области
Локальный$_4$	Непроницаемый для прочих операций	Оценка вычислительной эффективности языка

Рис. 1.

Значение Интерпретация Типичный контекст применения Локальный$_1$ Циклический Алгоритмизация разных правил/актов применения одного правила Локальный$_2$ Ингерентный Добавление новых факторов, влияющих на уже построенную структуру Локальный$_3$ Контекстно-свободный Сужение/расширение синтаксической области Локальный$_4$ Непроницаемый для прочих операций Оценка вычислительной эффективности языка

Все четыре употребления предполагают контраст локальных и нелокальных правил/факторов и формальный аппарат, позволяющий исследовать явления, лежащие за рамками Контекстно-Свободных грамматик. Поскольку данные четыре употребления в лингвистической литературе связаны между собой подобием семантической сети, а совместное применение локальных/нелокальных правил в любом из этих четырех употреблений

[8] 'this must be a non-local factor, and not the structure of the pronoun itself, (выделение наше – *А.Ц.*)– contrary to what Cardinaletti and Starke hypothesize. … the restriction on coordinating nonhuman pronouns may be due to a particular instance of a more general pragmatic principle.'.

[9] Это нельзя сделать внутри формализма И.А.Мельчука, так как он отвергает представление об изначально линейно-упорядоченном дереве предложения, см. обсуждение в (Циммерлинг 2002, 186-188).

085-Cimmerling_A:_Layout 1 18.05.2008 21:09 Page 554

возможно лишь на основе грамматик, мощность которых достаточна для описания любого естественного языка, мы будем считать, что термины «локальный» и «глобальный» покрывают все четыре области значения, и позволим себе в заключительной части статьи, где обсуждаются факты русского языка, использовать их как интуитивно ясные, не уточняя, которое из четырех значений более всего подходит в конкретном случае.

Современный русский язык оставляет говорящему богатые возможности расположить элементы предложения в разном порядке и выбрать слово-носитель главного фразового акцента (Ковтунова 1976, 6-10; Иванова-Лукьянова 2004, 16-17). Это особенно заметно при сопоставлении с такими языками, как немецкий, английский или датский, где многие линейные порядки исключены из-за нормативных запретов, а выбор носителя акцента в группах типа прилагательное + существительное или глагол + дополнение предопределяется структурой самой фразы, т.е. локально. Возникает соблазн заключить, что в языках типа русского порядок слов и размещение фразового акцента целиком зависят от произвола говорящего, т.е., в терминах данной статьи, определяются глобальными правилами в отсутствие локальных. При этом, однако, остается загадкой, как носители русского языка общаются, дешифруя намерения друг друга: если же за комбинаторикой линейных порядков и интонационных схем не стоит коммуникативно значимой информации, непонятно, зачем язык допускает это многообразие.

Мы предполагаем, что естественный язык не может функционировать подобным образом, и что локальные правила размещения фразового акцента и порядка слов существуют, несмотря на то (или благодаря тому что – *А.Ц.*) они слабо рефлектируются говорящими. Ориентиром для нас будет комбинаторная модель фразовой просодии, опирающаяся на книгу И.И.Ковтуновой (Ковтунова 1976) и разработанная в трудах Н.Д.Светозаровой, С.Д.Кодзасова и Т.Е.Янко (Светозарова 1993; Кодзасов 1996; Кодзасов 1996а; Янко 1991; Янко 2001; Янко 2007). Возьмем за основу перцептивный, а не экспериментально-фонетический подход к интонационным явлениям и примем без доказательств, что носители русского языка способны однозначно определить место акцента в любой фразе: в противном случае нам пришлось бы исследовать фонетические механизмы, стоящие за фразовыми акцентами, и, в частности, решать вопрос, можно ли описывать их локально, выявляя уникальные черты ударного слога/слова и абстрагируясь от окружения, ср. подход Дж. Пьерхамберт (Pierrehumbert 1980), или же необходимо прибегать к глобальным правилам приписывания акцентов и просматривать всю просодическую область, на чем основан подход Н. Грённум (Grønnum 1992, 48-50). Примем также без доказательства, что просодически значимые фразовые акценты в русском языке всегда имеют тональный компонент, независимо от наличия других составляющих (амплитуды, специальных гортанных настроек и т.д.). Для процедуры определения слова-акцентоносителя сделанное допущение о (преимущественно) тональном характере русского ударения избыточно, но оно полезно на дальнейшей стадии анализа, если исследуется трансформационные отношения, связывающие между собой предложения с уже приписанными фразовыми акцентами.

В наиболее детализованной форме алгоритм выбора акцентоносителя описан в (Янко 2007). Согласно Т.Е.Янко, в русском языке, если не действуют возмущающие факторы, о которых ниже, соблюдается иерархия акцентоносителей, зеркально отражающая иерархию членов предложения. Чем ниже грамматический статус актанта многоместного предиката, тем выше вероятность, что именно этот актант получит фразовый акцент[10]:

(i) *Предикат* (P) > *сирконстанты* (C) > -*актанты* (A), в порядке, заданном актантной структурой предиката (- A1, A2, A3…A_n).

Тестом для проверки служат предложения с т.н. неингерентной темой, где ни один из элементов не выделяется семантически:

(3а) <Что случилось? Чем ты так взволнован?> Вася Марго ↘МАНТО подарил[11].

(3б) <Что случилось? Чем ты так взволнован?> Вася ↘↘МАНТО Марго подарил.

Иерархия (i) полезна тем, что позволяет проверить грамматический статус выражения как актанта/сирконстанта, прямого/косвенного дополнения там, где это интуитивно неочевидно. Примеры (3аб) показывает, в

[10] Выведенная Т.Е.Янко формулировка алгоритма учитывает 9 факторов: (a) синтаксическая структура (NP, VP, S, две или более группы); (b) активация (имя референта, названного в предтексте, исключается из списка претендентов на роль акцентоносителя ремы; (c) актантная структура предиката; (d) набор сирконстантов; (e) синтаксическая структура в терминах членов предложения; (g) структурная схема предложения; (h) идиоматичность заполнения валентностей; (i) активность актантов; (j) внутренняя структура именных, глагольных и нексусных групп, которые могут представлять собой атрибутивные и сочиненные группы и в которых действуют внутренние правила выбора акцентоносителя (Янко 2007). Но такое количество факторов нужно только в том случае, если синтаксический анализ производится одновременно с выбором акцентоносителя. Если же модуль, отвечающий за расстановку акцентоносителя, получает на входе уже готовую синтаксическую структуру, факторы (a), (c-g), (i-j) удобно рассматривать вместе. Фактор (h) – идиоматичность актантов – возможно задать на словаря, в зоне «модель управления предиката». Проблемным выглядит только фактор (b) – активированности/неактивированность референта в предтексте, так как он, в отличие от всех предыдущих, требует обращения к контексту шире отдельного предложения.

[11] Здесь и далее маркировки тональных акцентов записываются слева от словоформы. Запись '↘' символизирует нисходящий акцент ИК 1, который в повествовательном предложении маркирует Рему. Запись '↗' символизирует восходящий

085-Cimmerling_A:_Layout 1 18.05.2008 21:09 Page 555

частности, что у русских глаголов класса «давать», «дарить», именно актант в дат.п ., имеющий семантику адресата действия, следует считать прямым дополнением, т.е. акцентным приоритетом обладает не он, а актант в вин.п., с семантикой пациенса, см. обсуждение в (Циммерлинг 2007).

Высказывания (3в) и (3г), где носителем фразового акцента будет дополнение в дат.п., возможны, но они, в отличие от (3аб), не нейтральны семантически, и выделенный элемент в них получает значение контрастной ремы:

(3в) Вася манто ↘МАРГО подарил.

(3г) Вася ↘↘МАРГО манто подарил.

Варианты (3вг) будут оправданы либо в ситуации, где говорящий подчеркивает, что подарил манто именно X-у, а не Y-у, либо в ситуации, где о манто уже шла речь. Последний случай подпадает под правило, которое Янко интерпретирует как часть базового алгоритма расстановки фразового акцента.

(ii) актант исключается из списка кандидатов, если его референт уже активирован в дискурсе.

По нашему мнению, (ii) можно задать и как контрправило к (i), и как частный случай последнего. В любом случае, (ii) заставляет лингвистический процессор просматривать фрагмент больше элементарного предложения, в то время как для других пунктов алгоритма Янко этого можно избежать. Если разрешить алгоритму Янко просматривать/конструировать левый контекст в парах предложений (не обязательно смежных в тексте), размещение акцента на глаголе в примере (4) можно будет предсказать.

(4) <Вася утром ходил к врачу>… ↘Осмотрел $_0$врач $_0$Васю и <не нашел у него плеврита>.

Если (ii) не считается составной частью (i), (i) будет локальным правилом. Если же (ii) считать составной частью (i), то статус (i) зависит от того, разрешим ли мы алгоритму Янко просматривать фрагменты текста больше одного предложения. Других препятствий для определения акцентной иерархии (i) как локального правила нет: она автоматически реализуется как в предложениях с неингерентной темой, так и в предложениях с внешне выраженной темой. Точно также, для локуса акцента нерелевантно, является ли он нисходящим (запись '↘', '↘↘', как в (3аг) и (4), или же восходящим ('↗'), как в вопросительных предложениях (5аб) и (6).

(5а) <Что случилось? Чем ты так взволнован?> Вася Марго ↗МАНТО подарил?

(5б) <Что случилось? Чем ты так взволнован?> Вася ↗МАНТО Марго подарил?

(6) <Вася утром ходил к врачу>… ↗Осмотрел $_0$врач $_0$Васю?

На базовое локальное/квазилокальное$_2$ правило выбора акцентоносителя в русском языке наслаиваются два контрправила, нарушающие иерархию (i). Одно из них, впервые описанное Н.Д.Светозаровой, состоит в том, что в некоторых речевых ситуациях или мини-жанрах текста (ср. обращение, перечисление, зачитывание объявления) акцент смещается на край атрибутивной или сочиненной составляющей (обычно – на левый край). Тем же можно объяснить контраст между предложениями (7) и (8), произнесенными одним и тем же информантом.

(7) Мою подругу зовут Полина ↘Игоревна. (базовое правило).

(8) ↘Полина Игоревна, ну как же так! (обращение с упреком).

Правило, заставляющее смещать акцент на левый край составляющей в примере (8), носит глобальный характер, его нельзя задать как локальное: говорящий/лингвистический процессор должен уметь распознать, что пример (8) произносен в контексте обращения с упреком.

Второе контрправило, несводимое к (i), описано в (Янко 2007). Оно состоит в том, что при некоторых маркированных иллокутивных актах (настойчивые просьбы, заискивающие просьбы, мечты, воспоминания, недоумения, идентификация), акцент внутри составляющей смещается. Ср. контраст между (9) и (10).

(9) Дайте мне [бланки на ↘**визу**] (нейтральная просьба посетителя, соблюдающего дистанцию с адресатом).

(10) Мне бы [↘**бланки** на визу] (говорящий заискивает перед адресатом).

С утверждением (11а), где содержится утверждение о незнании некоторого положения дел, контрастируют примеры (11бв), где говорящий выражает свои эмоции.

(11а) Не знаю,[куда запропастились мои ↘**очки**]. (нейтральное утверждение о незнании р).

(11б) Не знаю, [↱**куда-а** запропастились мои очки]. (выражение эмоций).

(11в) [↱**Куда-а** мои очки запропастились? (то же)[12].

Глобальные контрправила, связанные со смещением акцентоносителя на край составляющей или с действием особых иллокутивных сил, изучены в деталях хуже, чем базовые правила (i-ii), хотя отклонения от стандартного принципа размещения фразового акцента бросаются в глаза: некоторые фонетисты даже говорят в этой связи о новых тенденциях в развитии русской разговорной речи (Светозарова 1993, 197). По нашему мнению, здесь нет парадокса, так как для точного описания упомянутых контрправил требуется сопоставление с базовым

акцент ИК 3, который в повествовательном предложении маркирует Тему. Запись '↘↘' символизирует нисходящий акцент, сопровождающийся динамическим усилением (ИК 2): в повествовательном предложении данный акцент маркирует эмфатическую Рему. Запись '$_0$ X' символизирует отсутствие тонального акцента, что отражает понижение статуса элемента в коммуникативной иерархии. О прочих акцентных маркировках см. выше в тексте.

[12] Примеры Т.Е.Янко.

085-Cimmerling_A:_Layout 1 18.05.2008 21:09 Page 556

правилом, которое осознается говорящими слабо. Это тоже не кажется нам парадоксальным. Главная причина именно в высокой степени формализации базового правила: как указано, выше иерархия (i) зеркально отображает грамматическую иерархию ИГ, изучение которой нельзя отнести к главным достижениям академической русистики[13]. Немаловажно и то, что локальное правило (i), как и его расширение, квазилокальное$_2$ правило (ii), не действует изолированно: в соответствие с предсказаниями, сделанными в первой части статьи, на него наслаиваются глобальные контрправила, нарушающие стройность исходного распределения.

Последний раздел статьи посвящен т.н. Линейно-Акцентным преобразованиям, соединяющим между собой высказывания с уже приписанными коммуникативными статусами Темы и Ремы. Примем без доказательства тезис о том, что основным средством выражения актуального членения в русском языке являются порядок слов и интонация, точнее – комбинации линейных порядков с комбинацией фразовых акцентов (Ковтунова 1976, 8-10). В таком случае, для того, чтобы создать требуемое исчисление, нам нужны три вещи: а) модуль, строящий дерево предложения[14]; б) алгоритм выбора акцентоносителя в составляющих (мы уже знаем, что для русского языка он задается локальным правилом (i); в) тональный алфавит, т.е. список релевантных тональных акцентов. Последние названы в работах Г.Н.Ивановой-Лукьяновой и О.Йокоямы «интонемами» или «тонемами» (Иванова-Лукьянова 2004, 6-13; Yokoyama 2001, 4) и анализируются как двусторонние сущности, т.е. интонационные контуры, соотнесенные с некоторым характерным для них коммуникативным содержанием, ср. (Николаева 1982) . Из семи интонационных контуров, выделяемых в русском языке Е.А.Брызгуновой (Грамматика 1980 I, 97-122) и ее последователями, синтаксически релевантны, по нашему убеждению, четыре: ИК 1, ИК 2, ИК 3 и ИК 6. Прочие – ИК 4, ИК 5 и ИК 7– принадлежат «акцентному словарю», а не «акцентной грамматике», так как они соотнесены со специальными типами речевых актов и не задействованы в преобразованиях порядка слов. Синтаксически релевантные акценты отличаются от прочих русских «тонем» еще тем, что у каждого из них обнаруживается ровно две разных коммуникативных функции, в зависимости от того, в высказывании какого типа они представлены: ИК 1 и ИК 2 в повествовательном предложении маркируют рему, а в вопросах – несобственно вопросительный компонент[15], ИК 3 в повествовательном предложении является маркером ремы, а в вопросах – маркером вопросительного компонента, ИК 6 может маркировать как элемент со статусом акцентно ослабленной ремы, так и актант со статусом дислоцированного компонента ремы. Ничего подобного в употреблении «словарных» тонем ИК 4, ИК 5 и ИК 7 не прослеживается. Поскольку мы трактуем русские ИК как двусторонние знаки, а не просто как контурные просодии, мы отвергаем тезис о существовании каких-либо промежуточных форм вроде *ИК$^{1-4}$, *ИК$^{1-2}$, *ИК $^{2-3}$ и т.п.[16] Промежуточных форм между ИК 1 и ИК 4, ИК 2 и ИК 3 ни в нашем формализме, ни, как мы полагаем, в системе русского языка, быть не может, поскольку соответствующие интонационные контуры выражают разное коммуникативное содержание. Что касается двух синтаксически релевантных нисходящих акцентов, ИК 1 и ИК 2, то они соотносятся между собой как нейтральный vs усиленный варианты акцента ремы/несобственно вопросительного компонента, противопоставление их поддерживается регулярными линейно-акцентными преобразовании. Все исследователи признают, что ИК 2 характеризуется повышением интенсивности (динамическим акцентом в терминах С.В.Кодзасова), на ударном слоге, но мы хотели бы подчеркнуть и наличие тональной составляющей, а именно, опережающего движения тона (раннего тайминга) при ИК 1 сравнительно с ИК 2. Сходным образом, два синтаксически релевантных восходящих акцента, ИК 3 и ИК 6, фонетически отличаются не только наличием/отсутствием спада на заударных и разной амплитудой подъема на ударном слоге, но и разным таймингом – поздним для ИК 6 и ранним для ИК 3.

---

[13] В русистике со времен А.М. Пешковского существует традиция сводить понятие дополнения к предположительно более ясным понятиям «управляемого второстепенного члена» и «переходности» (Пешковский 1938, 267-269). В русле этой же традиции написана последняя академическая грамматика русского языка, где понятие «дополнение» не упомянуто вовсе: в предметном указателе к тому «Синтаксис», а также в названиях разделов этого тома данный термин не встретился ни разу (Грамматика 1980 II, 663-709).

[14] Является ли оно линейно упорядоченным заранее, или нет, мало не влияет на дальнейшие рассуждения.

[15] 'Несобственно вопросительный компонент' (= 'неконституирующий компонент вопроса') – составляющая, не являющаяся носителем иллокутивного значения вопроса и получающую положительную акцентную маркировку, отличную от акцентной маркировки вопросительного компонента (Янко 2001, 50). Так, в вопросе ↘*Статью* $_0$*Вася* ↗*вчера* $_0$*сдал?* собственно вопросительный компонент *Вася* маркируется восходящим акцентом '↗', а несобственно вопросительный компонент *статью* маркируется нисходящим акцентом '↘'.

[16] Ср. утверждение о наличии гибридных форм ИК в (Иванова-Лукьянова 2004, 10). Явно ошибочно утверждение О. Йокоямы, будто у ИК 1 «есть аллофон с легким подъемом на заударных слогах», т.е. ИК 4 (Yokoyama 2001, 4). В действительности, отдельные носители русского языка могут произнести ИК 4 в позиции, где нормативной интонацией является ИК 1, ср. рус. ?*Мне это* ↘↗*известно* вместо рус. *Мне это известно*, но отсюда не еще следует, что даже в некодифицированной речи тех лиц, которые позволяют себе такую замену, ИК 1 и ИК 4 смешиваются.

085-Cimmerling_A:_Layout 1 18.05.2008 21:09 Page 557

Интонационный Контур	Символическая запись	Автосегментная транскрипция	Коммуникативные Функции	Аллофоны
ИК 3	↗	LH*L- (восходящий акцент с ранним таймингом)	1) Тема. 2) Вопросительный компонент вопроса.	ИК 3 Q ИК 3 Contr
ИК 6	↱	L*HH- (восходящий акцент с поздним таймингом)	1) 2-я, Акцентно подавленная тема. 2) Левый компонент дислоцированной ремы.	
ИК 1	↘	HL*L- (нисходящий акцент с ранним таймингом)	1) Рема. 2) Невопросительный компонент вопроса.	ИК 1 Contr
ИК 2	↘↘	H*LL- (нисходящий акцент с поздним таймингом)	1) Рема 2) Невопросительный компонент вопроса.	

Рис. 2. Инвентарь релевантных тональных акцентов.

Вторая колонка таблицы содержит символическую акцентную маркировку, которая в настоящей статье записывается не справа (как у ряда наших предшественников), а слева от слова-акцентоносителя. Для того, чтобы построить исчисление линейно акцентных преобразований, к приведенному списку необходимо добавить еще одну маркировку, которой не соответствует фонологически значимая контурная просодия. Данная маркировка записывается в виде нижнего индекса ‘0’ слева от коммуникативной составляющей, которая понижается в коммуникативной иерархии в ходе линейно-акцентных преобразований и атонируется: ср. ↗ Дед посадил ↘репку. ⇒ ↱Посадил $_0$дед репку.

Интонационный Контур	Символическая запись	Автосегментная транскрипция	Коммуникативные Функции	Аллофоны
-	$_0$X	- (ровный тон/отсутствие акцента)	Снятый фонологический акцент	-

Рис. 3

Поскольку мы предлагаем не просодическую транскрипцию как таковую, а набор маркировок для регулярных синтаксически значимых преобразований, влияющих на акцентуацию составляющих, представляется неправильным включать в него избыточную для синтаксического процессора информацию о просодии фразы, например, сведения о уровне входящего тона (entering tone), амплитуде и глубине падения на заударных, наличии гортанной смычки или преаспирации и т.д.[17]

Линейные комбинации синтаксически релевантных акцентных маркировок несут в себе информацию, достаточную если не для реконструкции всей структуры составляющих, то для установления векторных связей между парами или *n*-ками предложений с общей лексико-синтаксической структурой, т.е. набором синтаксических позиций (same numeration) и их лексическим наполнением. В работе И.И.Ковтуновой множество ЛА-акцентных преобразований было названо *К* (оммуникативной)-парадигмой предложения (Ковтунова 1976, 34-58).

Определение.

Линейно-акцентные преобразования (ЛА-преобразования) определяются как трансформационные правила, одновременно меняющие порядок слов и акцентуацию минимум одной коммуникативной составляющей во фразе.

ЛА-преобразования не создают новых синтаксических позиций. Инвариантом ЛА-преобразований является лексико-синтаксическая структура предложения, остающаяся неизменной. ЛА-преобразования, в отличие первоначальных представлений об их природе в (Падучева 1984), являются несинонимическими в том плане, что они способны менять границы коммуникативных составляющих.

• Смена акцентной маркировки с ↗X ~ ↘X ~↘ ↘X на $_0$X трактуется как преобразование.

ЛА-преобразования являются КЗ-правилами. Возможность представить их в виде Мягко-Контекстно-Зависимых (Mildly Context-Sensitive) зависит от избранного формализма и от состава правил в конкретном языке.

[17] Все эти параметры, разумеется, важны для анализа языкового сигнала как такового и правомерно учитываются в современных работах по просодии фразы, ср. труды С.Оде и С.В.Кодзасова.

085-Cimmerling_A:_Layout 1 18.05.2008 21:09 Page 558

Модель ЛА-преобразований была впервые формализована Е.В.Падучевой (Падучева 1984), ср. (Падучева 2008, 107-119) и детально исследована применительно к русскому языку в (Янко 2001, 117-229). Опыт применения аппарата ЛА-преобразований к другим языкам см. в (Циммерлинг 2002, 190-208; 353-367). В (Zimmerling 2007; Zimmerling 2008) сделана попытка совместить идею ЛА-преобразований с моделью формально-синтаксического перемещения (Movement).

Как и наши предшественники, мы выделяем на множестве ЛА-преобразований предложения исходный нейтральный член, «у которого вклад коммуникативной структуры в семантическую минимален»[18]. Исходный вариант множества ЛА-преобразований повествовательного предложения имеет бинарную структуру (iii).

(iii) Topic + Focus (тема целиком предшествует реме, разрывы коммуникативных составляющих отсутствуют[19]).

И.И.Ковтунова, Е.В.Падучева, Т.Е.Янко исходят из того, что коммуникативное и формально-синтаксическое членение фразы автономны друг от друга, и что сказуемое или его часть не всегда являются ремой. Это положение безусловно верно (и даже тривиально) в качестве общетеоретического постулата, но отсюда не следует, что оно оправдывает себя при формализации ЛА-преобразований. Главными минусами такого подхода является то, что у многих предложений в языках со свободным порядком слов будет более одного исходного ЛА-варианта, ср. (12а) и (12б)

(12а). [$_{\mathbf{T}}$Профессор [↗Иванов]] [$_{\mathbf{F}}$в июне посетил [нашу ↘лабораторию]].

(12б) [$_{\mathbf{T}}$Нашу [↗лабораторию]] [$_{\mathbf{F}}$в июне посетил [профессор ↘Иванов]].

Кроме того, если не вводить постулат о базовом порядке слов, и не выбирать исходный ЛА-вариант по тому или иному формальному критерию, все ЛА-варианты могут оказаться непосредственно выводимы друг из друга, что явно неприемлемо. Если же добавить такой постулат и выводить (12б) из (12а) или наоборот, есть надежда, что алгоритм линеаризации будет работать эффективно, а множество ЛА-преобразований одного предложения удастся представить в виде конечной цепочки без замыканий и возвращений в исходную точку, т.е. в виде (несинтаксического) дерева.

Дополнительные плюсы привязки коммуникативной структуры к субъектно-предикатному членению в исходной последовательности видятся нам в том, что при таком подходе можно применять модель синтаксических перемещений и разграничить активно передвигающиеся элементы и сопутствующие изменения линейного порядка и акцентуации, которые в теории формальных грамматик могут быть сведены к рубрике “остаточное перемещение” (Remnant Movement)[20]. Так, в производном предложении рус. ↦*Посадил*$_{i}$ $_{0}$*дед* t$_{i}$ ↘*репку* активно перемещающимся элементом следует считать глагол *посадил*, который смещается в начало фразы и получает положительную акцентную маркировку ‘↦’, сообразно изменению своего коммуникативного статуса по сравнению с исходным предложением: $_{0}$X / ↦X’. Напротив, атонирование ИГ *дед*, которая попадает в поствербальную позицию в силу перемещения глагола влево, в начало предложения, интерпретируется нами как сопутствующее остаточное перемещение: ↗X /$_{0}$X .

Мы принимаем критерий (iv), по которому в базовом ЛА-варианте коммуникативное членение должно гармонизироваться с бинарным членением на подлежащее[21] и сказуемое.

(iv) В базовом ЛА-варианте повествовательного предложения границы тематической составляющей и группы подлежащего совпадают. Группа сказуемого в базовом ЛА-варианте всегда рематична, в ее составе выделяется синтаксическая группа носителя рематического акцента (Focus Proper) и переходная зона между собственно темой и носителем рематического акцента (Transition).

	Topic		Focus	
Communicative structure		Topic proper	Transition	Focus proper
Syntactic structure	Grammatical subject		Grammatical Predicate	
	External argument		Verbal head	Complements (Internal arguments & adjuncts)

Рис. 4 Прототипическое соотношение коммуникативной и синтаксической структуры для глагола с внутренними фразовыми зависимыми.

[18]. Та же идея используется в (Янко 2001, 137).

[19]. Данное положение традиционно, см. (Ковтунова 1976, 15).

[20]. В древо-присоединяющих грамматиках Remnant Movement понимается как перемещение (под) дерева, в составе которого уже произошло перемещение. Remnant Movement описывается Мягко-контекстно-зависимыми языками (Stabler 1999; Gärtner & Michaelis 2007).

[21]. В первом приближении можно ограничиться случаями со стандартным ненулевым подлежащим в им.п. Нулевые подлежащие и случаи эллипсиса подлежащего в данном контексте следует безусловно исключить из рассмотрения, а статус т.н. косвенных подлежащих в дат.п/вин.п. в рамках ЛА-преобразований требует дополнительной проверки.

085-Cimmerling_A:_Layout 1 18.05.2008 21:09 Page 559

При необходимости установить не тернарное, а бинарное коммуникативное членение переходная зона (Transition) учитывается в составе ремы, а не темы[22]. В отличие от отечественных русистов мы анализируем носитель главного акцента (Focus Proper) не как отдельную словоформу-акцентоноситель, но как группу, которая целиком подвергается перемещениям и прочим синтаксическим операциям. В отличие от подхода в (Янко 2001, 180, 198), мы отказывается интерпретировать атонирование, т.е. помещение составляющих, не являющихся клитиками, в слабую фразовую позицию, как движущий фактор, создающий те или иные синтаксические конфигурации, и видим в атонировании сопутствующий феномен (Remnant Movement), вызванным перемещением других элементов. В этой связи мы принимаем постулат (v):

(v). В ЛА-преобразованиях русского языка возможно выделить основное (Active Movement) и сопутствующее (Remnant Movement) преобразование. Основное преобразование связано с перемещением коммуникативных составляющих в конечные позиции, где они получают положительную акцентную маркировку (‘↗X’, или ‘↘X’, или ‘↘↘X’, или ‘↦X’). Преобразования, при которых попадают в конечные позиции с нулевой акцентной маркировкой $_0$X, всегда являются сопутствующими. Возможны также сопутствующие преобразования, при которых составляющие попадают в конечные позиции, где они получают положительную акцентную маркировку.

Для облегчения выбора исходного ЛА-варианта необходимо принять еще два постулата, идущие вразрез с практикой лингвистического описания.

(vi). ЛА-варианты повествовательных предложений, предполагающие коммуникативную расчлененность на тему и рему, обладают приоритетом над коммуникативно нерасчлененными (thetic) предложениями. Последние в формальном плане всегда являются производными от первых, но не наоборот.[23]

(vii) ЛА-варианты повествовательных предложений, где некоторая составляющая имеет положительную акцентную маркировку, при прочих равных условиях, обладают приоритетом над ЛА-вариантами, где та же составляющая имеет нулевую маркировку.[24]

Минусом постулатов (vi) и (vii) является то, что они в некоторых случаях автоматически вынуждают признать в качестве исходных такие ЛА-варианты, которые являются более редкими и менее естественными, чем ЛА-варианты, сигнализирующие коммуникативно нерасчлененные значения, ср. (13а) и (13бв), (14а) и (14б).
(13а) [[$_{\mathbf{T}}$↗Весна] [$_{\mathbf{F}}$ ↘пришла] ⇒ (13б) [$_{\mathbf{F}}$ t$_i$ $_0$Пришла [$_{\mathbf{F}}$ ↘весна$_i$]] ⇒ (13в) [$_{\mathbf{F}}$ t$_i$ [$_{\mathbf{F}}$↘↘весна]$_j$ $_0$пришла [$_{\mathbf{F}}$ ↘t]$_j$].
(14а) [[$_{\mathbf{T}}$↗ Скамейки] [$_{\mathbf{F}}$ ↘установят]. (ответ чиновника на запрос о скамейках)
(14б) [$_0$Установят [↘скамейки]], <и сквер обретет прежний вид >. (нерасчлененное сообщение о событии *p*).
(14в) [$_{\mathbf{F}}$ t$_i$ $_0$Установят [$_{\mathbf{F}}$ ↘скамейки] $_i$].

Плюсом избранного формализма является то, что он, как отмечено выше, позволяет задать ориентированный граф из ЛА-вариантов, без замыканий, и не выводить варианты друг из друга в хаотичном порядке. Из записи примеров (13ав) видно, в частности, что (13в) ↘↘*весна* $_0$*пришла* нельзя получить непосредственно из (13а) ↗*Весна*↘*пришла*: вначале нужно инвертировать ИГ *весна*, переместив ее правее глагола *пришла*, который в результате данного перемещения подвергается сопутствующему преобразованию – атонируется. И лишь потом можно прибегнуть к ‘обратному’ перемещению – в конечную позицию левее глагольного элемента, где она получает усиленный рематический акцент ‘↘↘’ (ИК 2). Этот этап деривации вполне соответствует интуитивному ощущению, поскольку ИК 2 в русском языке регулярно возникает именно при регрессивном перемещении элемента в пределах глагольной группы. Однако интуитивное (не теоретическое) оправдание первичности (13а) ↗*Весна* ↘*пришла* по сравнению с (13б) $_0$*Пришла* $_{\mathbf{F}}$ ↘весна$_i$ столь легко подыскать не удается.

Полный список ЛА-преобразований, выделенных нами в русском языке, составляет 7-8 операций[25].

[22]. В исходном ЛА-варианте коммуникативно переходный элемент (Transition) в русском языке по определению имеет нулевую акцентную маркировку, однако он может перемещаться в конечные позиции с положительной маркировкой. Для других языков, где глаголы и связки не атонируются автоматически, требуется иное описание.

[23]. Ср. например, теории Т.Л.Кинг и М.Бабёнышев: оба автора полагают, что нейтральным порядком слов (unmarked word order) является, такой, где нет ни «топикализации», ни «фокализации» (no Topic and Focus Movement takes place), что возможно лишь в «дискурсивно-нейтральных» высказываниях (discourse-neutral sentences), отвечающих на вопрос «что случилось». (King 1995; Babyonyshev 1996, 18). При этом Т.Л.Кинг постулирует для русского языка базовый порядок #VSO, так как якобы только такие высказывания могут быть коммуникативно нерасчлененными; порядки SVO, OVS она признает коммуникативно расчлененными, возникающими в результате перемещения в предфинитную позицию элементов со статусом Темы либо Ремы. Такое описание идейно, но неправильно: в русском языке коммуникативная нерасчлененность может быть соотнесена как с порядком VS, так и с порядком SV. М.Бабёнышев исходит из того, что предфинитная позиция в русском языке всегда должна быть заполнена, причем ровно одной синтаксической категорий (т.н. параметр EPP): эта категория, с известными оговорками, интерпретируется ею как ‘Подлежащее’ (ibid., 34-50). Такое описание тоже нельзя считать удовлетворительным.

[24]. Постулат (vii) равнозначен требованию о том, что положительная акцентная маркировка Тем и Рем является встроенной характеристикой предложения и не может добавляться или меняться ad hoc, в то время как атонирование является результатом ЛА-преобразований.h

[25]. Подробнее см. работы (Zimmerling 2007), (Zimmerling 2008).

085-Cimmerling_A:_Layout 1 18.05.2008 21:09 Page 560

Формат статьи не позволяет обсудить их подробно, поэтому ограничимся их перечислением и представим символическую запись. Используются четыре основных символа для коммуникативной составляющих – F (Рема), T (Тема) и Tr (Переход), SF (Дислоцированный компонент Ремы); дополнительные символы вводятся, если данное преобразование селективно может реализоваться лишь при строго определенной синтаксической характеристике элемента, см., например, запись Head (Вершина Группы) ниже в п. 7. Знак '↠' означает 'перемещение элемента правее исходной позиции', знак '↞' – 'перемещение элемента левее исходной позиции', знак '↓' – 'атонирование элемента'. Запись вида 'X/F ↠ ' читается 'перемещение элемента X вправо в конечную позицию Ремы', запись вида '↞ X/T' читается 'перемещение элемента X влево в конечную позицию Темы', запись вида '↞ Tr/T' читается 'перемещение элемента из исходной позиции Tr (Переход) влево в конечную позицию Темы' и т.д. Связка '&' ставится между записью основного и сопутствующего преобразования (ср. X/T & Y/F), выражая детерминистскую связь между ними, в типичном случае – вследствие того, что активно перемещающийся элемент X пересекает на своем пути элемент Y, подвергающийся сопутствующему преобразованию. Таким образом, запись **X/F ↠ & ↓ T**, соответствующая ЛА-преобразованию Right Focus Movement, полностью читается так: 'Продвижение элемента X вправо в конечную позицию Ремы, пересекающего на своем пути узел T, который вследствие этого атонируется'.

	Название	Операция	Основное (активное) перемещение	Сопутствующее изменение
1.	Right Focus Movement	X/F ↠ & ↓ T	X/F ↠	↓ T
2.	Left Focus Movement	↘F/↘↘F ↞ & Tr	↘F/↘↘F ↞	Tr
3.	Verb Topicalization	Tr/T ↞ & ↓T	Tr/T ↞	↓T
4.	Dislocation	Tr/SF ↞& ↓T	Tr/SF ↞	↓T
5.	Verb Focalization	Tr/F ↞ & ↓T , ↓F	Tr/F ↞	↓T, ↓F
6.	Topic-Focus Inversion	F/T ↞& T/F ↠	F/T ↞	T/F ↠
7.	Head Extraction	↞ Head/ F ~ T ~ SF & ↓ T	↞ Head/ F ~ T ~ SF	↓ T
8.	Focus Superposition	X/F ↠, ↘F/↘↘F	X/F ↠, ↘F/↘↘F	**

Рис. 7. Линейно-акцентные преобразования в русском языке.

Right Focus Movement (Перемещение элемента вправо в позицию конечной ремы).

(15) [$_{\mathbf{T}}$ ↗Моцарт] [$_{\mathbf{F}}$ [$_{\mathbf{FP}}$ ↘играет]], <а ↗скрипка ↘поет> ⇒ [$_{\mathbf{F}}$ $t_{i\ 0}$играет [$_{\mathbf{FP}}$ ↘Моцарт$_{i}$]].

См. также примеры (13б) и (14б). В других случаях Right Focus Movement может быть не основным, а сопутствующим перемещением (см. ниже).

Left Focus Movement (Перемещение элемента влево из позиции конечной ремы, в позицию в начале группы сказуемого, предшествующую вершине глагольной группы).

(16) [$_{\mathbf{F}}$ $t_{i\ 0}$играет [$_{\mathbf{F}}$ ↘Моцарт$_{i}$]] ⇒ [$_{\mathbf{F}}$ [$_{\mathbf{FP}}$ ↘↘МОЦАРТ$_{i}$]$_{j}$ $t_{i\ 0}$играет t_{j}].

Left Focus Movement действует только на предложения, которые порождаются при помощи Right Focus Movement. В русском языке это преобразование может применяться циклически, что формально доказывается в (Zimmerling 2008). Наличие/отсутствие перед группой сказуемого акцентно выраженной темы не играет роли, ср. (17б) и (17в).

(17а) [$_{\mathbf{T}}$ ↗Катя] [$_{\mathbf{F}}$родила своему мужу [$_{\mathbf{FP}}$ ↘мальчика]].

(17б) [$_{\mathbf{T}}$ ↗Катя] [$_{\mathbf{F}}$ [$_{\mathbf{FP}}$ ↘↘МАЛЬЧИКА]$_{j}$ [своему мужу]$_{i}$ родила t_{j} t_{i}].

(17в) [$_{\mathbf{F}}$ [$_{\mathbf{FP}}$ ↘↘МАЛЬЧИКА]$_{j\ 0}$Катя [своему мужу]$_{i}$ родила t_{j} t_{i}].

Verb Topicalization (Перемещение глагола влево в начало фразы в позицию основной Темы).

(18а) [$_{\mathbf{T}}$ ↗Вася] [$_{\mathbf{F}}$хорошо [$_{\mathbf{FP}}$↘объясняет] ⇒ (18б) # [$_{\mathbf{T}}$ ↗ Объясняет]$_{i\ 0}$Вася t_{i} [$_{\mathbf{F}}$ [$_{\mathbf{FP}}$↘хорошо]].

Dislocation (Дислокация ремы, с перемещением глагола влево в начало фразы в позицию Дислоцированного компонента).

(19а) [$_{\mathbf{T}}$ ↗Вася] [$_{\mathbf{F}}$обманул [$_{\mathbf{FP}}$↘покупателей] ⇒ (19б) # [$_{\mathbf{F}}$ [$_{\mathbf{SF}}$ ↦ Обманул]$_{i\ 0}$ Вася t_{i} [$_{\mathbf{FP}}$↘покупателей]] <и радуется>.

Verb Focalization (Перемещение глагола влево в начало фразы в позицию основной Ремы).

От двух предыдущих данное ЛА-преобразование отличается тем, что глагол попадает в позицию основной ремы, а все глагольные актанты атонируются.

(20а) [$_{\mathbf{T}}$ ↗Вася] [$_{\mathbf{F}}$обманул [$_{\mathbf{FP}}$ ↘девушку].

(20б) <Девушка доверилась Васе> # [[$_{\mathbf{FP}}$↘Обманул]$_{i\ 0}$Вася $t_{i\ 0}$девушку <что уж тут скрывать>.

Topic-Focus Inversion (Инверсия темы и ремы).

085-Cimmerling_A:_Layout 1 18.05.2008 21:09 Page 561

Данное ЛА-преобразование является одним их важнейших и включает два разнонаправленных перемещения в позиции, где они сохраняют положительную акцентную маркировку. Главной операцией является перемещение бывшей Ремы в позицию начальной Темы (в иных терминах – смена Темы). Выдвижение бывшей Темы в позицию конечной Ремы мы считаем сопутствующим процессом: так как предложение без ремы будет дефектным, выдвижение элементов в оказавшуюся вакантной позицию Ремы (Remnant Right Focus Movement) носит сугубо компенсаторный характер.

(21) [$_{\mathbf{T}}$ ↗Моцарт] [$_{\mathbf{FP}}$ ↘играет] ⇒ [[$_{\mathbf{T}}$ ↗играет]$_{i}$ t$_{j}$] [$_{\mathbf{F}}$ [$_{\mathbf{FP}}$ ↘Моцарт]$_{j}$ t$_{i}$].

(22) [$_{\mathbf{T}}$ Гуси и [$_{\mathbf{TP}}$ ↗лебеди]] [$_{\mathbf{F}}$ опротивели [$_{\mathbf{FP}}$ ↘Марусе] ⇒ [[$_{\mathbf{T}}$ опротивели [$_{\mathbf{TP}}$ ↗Марусе]$_{i}$ t$_{j}$] [$_{\mathbf{F}}$ Гуси и [$_{\mathbf{FP}}$ ↘лебеди] t$_{i}$].

(23) [$_{\mathbf{T}}$ ↗Котенок] [$_{\mathbf{F}}$ сидит [$_{\mathbf{FP}}$ ↘на шкафу]]⇒ [[$_{\mathbf{T}}$ ↗на шкафу]$_{i}$ t$_{j}$] [$_{\mathbf{F}}$ сидит [$_{\mathbf{FP}}$ ↘Котенок] t$_{i}$] .

Примеры (21-23) сильно отличаются по своей длине и синтаксическому составу. Это не должно затемнять их фундаментальную близость в плане ЛА-преобразования.

Head Extraction (Вынос синтаксической вершины группы и перемещение ее влево, в начало вышестоящей группы)[26].

Данное преобразование изучено пока недостаточно. Известно, что при выносе элемента в состав вышестоящей группы синтаксические вершины обнаруживают большую подвижность, чем зависимые элементы.

(24а) (↗) Я $_{0}$тебе [зашила [синюю ↘кофточку]].

(24б) # [[↘↘Кофточку]$_{i}$ $_{0}$я $_{0}$тебе зашила [t$_{i}$ синюю].

(24в) * # [[↘↘Синюю]$_{i}$ $_{0}$я $_{0}$тебе зашила [кофточку t$_{i}$].

(24г) ?? #[[↦ Синюю]$_{i}$ $_{0}$я $_{0}$тебе зашила [кофточку t$_{i}$].

При Head Extraction элементы, пересекаемые на своем пути перемещающейся влево вершиной, атонируются, что заставляет считать Head Extraction подлинным ЛА-преобразованием. Вместе с тем, в конечной позиции извлеченная синтаксическая вершина может иметь разные акцентные маркировки и получать разные коммуникативный статус. Возможно, имеет смысл выделять не одно, а несколько видов Head Extraction, аналогично тому, что предложено выше для перемещения глагола в начальную позицию.

Focus superposition (Суперпозиция ремы[27], порождение структуры, где одновременно имеются начальная и конечная рема).

Данное преобразование невозможно в живой речи – оно используется дикторами при зачитывании письменного текста, а также русскими журналистами, которые научились порождать аномальные письменные тексты, как бы рассчитанные на зачитывание вслух в программе новостей. Ср. пример (25)

(25) [$_{\mathbf{F}}$Сильный [$_{\mathbf{FP}}$ ↘↘взрыв]] прогремел сегодня в многоквартирном доме [$_{\mathbf{F}}$на севере британской [$_{\mathbf{FP}}$ ↘столицы]].

В формальном плане подобные примеры-монстры представляют собой ошибку деривации, когда автор текста одновременно прибегает к двум взаимоисключающим ЛА-преобразованиям – Right Focus Movement и Left Focus Movement.

Список литературы

1. M.Babyonyshev. 1996. Structural Connections in Syntax and Processing: Studies in Russian and Japanese. MIT.
2. A. Cardinaletti & M. Starke. 1999. The typology of structural deficiency // Clitics in the languages of Europe. Eurotyp 20-5. /Ed. by H. van Riemsdijk. Mouton de Gruyter. Berlin- New York. 429-467.
3. D. Ćavar & Ch. Wilder. 1999. 'Clitic third' in Croatian // Clitics in the languages of Europe. Eurotyp 20-5. /Ed. by H. van Riemsdijk. Mouton de Gruyter. Berlin- New York. 429-467.
4. N.Chomsky. 1977. On wh-movement. In P. Culicover, T.Wasow, and A. Akmajian, (eds.), Formal Syntax, pp. 71–132. Academic Press, New York.
5. N.Chomsky. 1993. A Minimalist Program for Linguistic Theory // The view from building 20. /Hale, K. S.L.Keyser (eds). Cambridge, Mass. MIT Press.
6. N.Chomsky. 2005. Three factors in language design. Linguistic Inquiry, 36: 1–22.
7. H.M.Gärtner & Jens Michaelis. 2007. Some remarks on the Locality Conditions and Minimalist Grammars, 162-195.
8. N. Grønnum. 1992. The Groundworks of Danish Intonation. University of Copenhagen.
9. T.L.King 1995. Configuring Topic and Focus in Russian. CSLI, Stanford.
10. A.K.Joshi. 1985. Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions? In D. R. Dowty, L.Karttunen, and A.M. Zwicky, (eds.), Natural Language Parsing. Psychological, Computational, and Theoretical Perspectives, 206–250. Cambridge University Press, New York.

[26]. О специфике операции Head Extraction с точки зрения формальной грамматики см. (Stabler 2001).

[27]. О термине см. (Янко 2001, 145).

085-Cimmerling_A:_Layout 1 18.05.2008 21:09 Page 562

11. J.B.Pierrehumbert. 1980. The Phonology of Emglish Intonation. M.I.T. Doctoral dissertation.

12. J.S.Ross. 1967. Constraints on Variables in Syntax. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

13. E. P.Stabler. 1997. Derivational minimalism. In Christian Retore, ed., Logical Aspects of Computational Linguistics. Springer, p. 68–95.

14. E. P.Stabler.1998. Acquiring languages with movement. // Syntax, 1: 72–97.

15. E. P.Stabler. 1999. Remnant movement and complexity. In G. Bouma, G.-J. M. Kruijff, E.Hinrichs, and R. T. Oehrle, (eds.), Constraints and Resources in Natural Language Syntax and Semantics, pp. 299–326. CSLI Publications, Stanford, CA.

16. E. P.Stabler. 2001. Recognizing head movement. In de Groote et al. (2001), pp. 245–260.

17. Y.Testelets. 2003. Are there Strong and Weak Pronouns in Russian? Formal Approaches to Slavic Linguistics #11: The Amherst Meeting 2002 / Browne, Wayles, Ji-Yung Kim, Barbara H. Partee, and Robert A. Rothstein (eds.)?. Michigan Slavic Publications. 515-538.

18. O.Yokoyama. 2001. Neutral and non-neutral intonation in Russian: A reinterpretation of the IK system //. Die Welt der Slaven. XLVI. 1-26.

19. A.Zimmerling. 2007. Topic-Focus Articulation, Verb Movement and the EPP in Russin // Slavic Linguistic Conference SLS 2. Berlin.

20.. A.Zimmerling. 2008. Locative Inversion and Right Focus Movement in Russian. Moscow.

Грамматика 1980. Русская грамматика. Т. 1, М., Наука. 1982.

21. Г.Н. Иванова-Лукьянова. 2004. Культура Устной Речи. Интонация, паузирование, логическое ударение, темп, ритм. М., Флинта. Наука.

22. С.В.Кодзасов. 1996. Комбинаторная модель фразовой просодии // Просодический строй русской речи. М.

23. С.В.Кодзасов. 1996а. Законы фразовой акцентуации // Просодический строй русской речи. М.

24. И.И. Ковтунова 1976. Современный русский язык: Порядок слов и актуальное членение предложения. М.

25. Т.М.Николаева. 1982. Семантика акцентного выделения. М.

26. С. Оде. 1995. Интонационная система русского языка в свете данных перцептивного анализа // Проблемы фонетики, Вып. II. М., 200-215. А.М.Пешковский. 1938. Русский синтаксис в научном освещении. М.

27. А.В.Павлова. 2007. Психолингвистический аспект инверсии // Вопросы психолингвистики. М., 2007, Т. 4, с. 73-80.

28. А.В.Павлова. 2007а. Роль акцентной структуры высказывания для перевода и лексикографии // Вестник ПГТК: проблемы языкознания и педагогики. № 10 (16). Пермь: ПГТУ.

29. Е.В.Падучева. 1984. Коммуникативная структура предложения и понятие коммуникативной парадигмы // НТИ, Сер. 2. N 10.

30. Е.В.Падучева. 2008. Высказывание и его соотнесенность с действительностью. 5-е изд. М.Ж ЛКИ..

31. Н.Д. Светозарова. 1993. Акцентно-ритмические инновации в русской спонтанной речи // Проблемы фонетики, Вып. 1. М., 189-198.

32. А.В. Циммерлинг. 2002. Типологический синтаксис скандинавских языков. М., «Языки славянской культуры», 896 стр.

33. А.В. Циммерлинг. 2007. Порядок слов в русском языке // Текст, Структура и семантика. Доклады XI международной конференции М.,МГГУ. 138-151.

34. Т.Е. Янко. 2001. Коммуникативные стратегии русской речи. М.: Языки славянской культуры.

35. Т.Е. Янко. 1991. Коммуникативная структура с неингерентной темой // Научно-техническая информация. Сер. 2, № 7.

36. Т.Е. Янко. 2007. Актантная структура как фактор фразовой просодии. Три принципа выбора акцентоносителя коммуникативно релевантного акцента // Типология языка и теория грамматики. Материалы международной конференции, посвященной 100-летию со дня рождения С.Д.Кацнельсона. СПб.

086-Cirulnik_L:_Layout 1 18.05.2008 16:26 Page 563

АЛГОРИТМ ИНТОНАЦИОННОЙ РАЗМЕТКИ ПОВЕСТВОВАТЕЛЬНЫХ ПРЕДЛОЖЕНИЙ ДЛЯ СИНТЕЗА РЕЧИ ПО ТЕКСТУ

ALGORITHM OF THE INTONATION MARKING OF NARRATIVE SENTECES FOR TTS SYNTHESIS

Цирульник Л.И. *(L.Tsirulnik@newman.bas-net.by),* ***Лобанов Б.М.*** *(Lobanov@newman.bas-net.by),*
Сизонов О.Г. *(Osizonov@yahoo.co.uk)*
Объединённый институт проблем информатики НАН Беларуси, Минск, Беларусь

Описывается алгоритм синтагматического членения и интонационной разметки повествовательных предложений, учитывающий позиционные и комбинаторные просодические факторы. Использование предложенного алгоритма при синтезе речи по тексту позволяет избежать так называемой «монотонности второго рода».

Введение

В [1] описаны общие принципы синтеза просодических характеристик речи по тексту, реализованные в системе «МультиФон». Первый из блоков подсистемы синтеза просодических характеристик речевого сигнала (просодический процессор), используя язык-зависимые ресурсы и правила, осуществляет анализ и просодическую разметку нормализованного орфографического текста. Анализ и просодическая разметка текста происходит в несколько этапов. На первом этапе осуществляется расстановка сильных и слабых словесных ударений, для чего используется грамматический словарь словоформ, содержащий пометы позиции ударения каждой словоформы, а также правила расстановки слабых и сильных словесных ударений, которые учитывают, в частности, принадлежность слова к знаменательным или служебным частям речи, его положение в предложении и ближайшее окружение. На следующем этапе – этапе объединения орфографических слов в фонетические слова и акцентные единицы (АЕ) – используются списки энклитиков и проклитиков, а также правила объединения в АЕ, которые также учитывают принадлежность «смежных» слов к определённым частям речи. На этапе разбиения текста на синтагмы и установки интонационного типа синтагм – завершающем этапе анализа и просодической разметки – используются правила синтагматического членения текста, согласно которым количество АЕ в синтагме не может превышать некоторого фиксированного количества (например, четырёх). Правила синтагматического членения и определения интонационных типов используют явные маркеры границ синтагм в тексте: знаки препинания, а также неявные, в частности, сочинительные и подчинительные союзы. Выходным данным блока анализа и просодической разметки является текст с пометами позиций ударения, границ фонетических слов и АЕ, а также синтагм с указанием интонационных типов каждой синтагмы.

Данная работа посвящена описанию алгоритма просодической разметки наиболее частотных компонентов входного текста – повествовательных предложений. На первом этапе осуществляется членение текста на предложения, и далее каждого предложения – на пунктуационные и лексические синтагмы. На втором этапе реализуется автоматическая маркировка интонационного типа каждой синтагмы.

1. Членение текста на предложения, пунктуационные и лексические синтагмы

Синтез речи осуществляется по предложениям, которые характеризуются достаточной степенью интонационной автономности в тексте и допускают наличие достаточно длительной паузы между ними (0,5 – 1,5 сек.).

Предложением считается отрезок текста, ограниченный знаками [.], [?], [?!], [!], [!!!]. Конец предложения может быть обозначен также знаком […], при условии, что следующее за ним слово начинается с большой буквы.

Предложением будем считать также заголовок всего текста или его части, в конце которого знак [.] может отсутствовать. Конец такого предложения обозначим знаком [*]. Кроме того, в отдельный тип выделяется предложение, ограниченное точкой в конце абзаца. Конец абзаца обозначим знаком [#].

Индикаторами членения предложения на пунктуационные синтагмы (ПС) являются знаки препинания.

086-Cirulnik_L:_Layout 1 18.05.2008 16:26 Page 564

Пунктуационными синтагмами будем считать предложение (при отсутствии в нём знаков препинания) или части предложения, ограниченные следующими знаками:

– точка с запятой [**;**],

– двоеточие [:],

– запятая [**,**],

– тире [–],

– открывающая скобка [**(**],

– закрывающая скобка [**)**],

– комбинация знаков [,–].

Таким образом, если предложение содержит *n* знаков препинания (включая знак конца предложения), то оно разбивается на *n* пунктуационных синтагм (*n* = 1,2,3,…). Определённым исключением из этого правила может служить ситуация, когда знак препинания стоит после сочинительного союза: *и, да, но и, так и, а, но, однако, зато, или, либо, то* и др. В этом случае предпочтительнее будет отказаться от установки синтагматической границы на месте этого знака препинания, хотя она и допустима для некоторого индивидуального стиля речи.

(1) Пример: "*Он быстро вошел и, увидя нас, внезапно остановился".*

Очевидно, что пунктуационные синтагмы могут быть различной длины (где под длиной понимается количество слов). Если длина синтагмы слишком большая (например, более 4-х слов), то следует убедиться, не содержит ли она некоторые простые лексические признаки (определённые слова или словосочетания), которые позволили бы разбить её на более мелкие лексические синтагмы (ЛС***)***. Экспериментальные исследования [2] показывают, что во многих случаях к таковым может быть отнесено присутствие следующих лексических признаков:

– соединительного союза «И».

(2) Пример[1]: «*Они посидели / и пошли гулять дальше».*

Раздел синтагмы – перед «И».

– разделительного союза «ИЛИ».

(3) Пример: «*Стоит ли нам сейчас пообедать / или подождать до 3-х часов*»? Раздел синтагмы – перед «ИЛИ».

– имён собственных (ИС).

(4) Пример: «*Сегодня певица Алла Пугачёва / решила выступить в нашем городе*». Раздел синтагмы – после последнего из следующих подряд ИС.

– аббревиатур (АБ).

(5) Пример: «*Возможность победы БНФ / вызывает большие сомнения*».

Раздел синтагмы – после АБ.

– названий разрядов чисел (РЧ).

(6) Пример: «*Два миллиона / десять тысяч / сто пять целых / двадцать пять сотых*».

Раздел синтагмы – после каждого РЧ.

– названий месяцев, слов «час, минута» при расшифровке даты и времени (ДВ).

(7) Пример: «*Десять часов / пять минут / десятого июня / седьмого года*».

Раздел синтагмы – после ДВ.

Указанный перечень не является полным и может быть расширен в процессе анализа всё более объёмных текстовых и речевых корпусов.

2. Маркировка интонационного типа синтагм

Категория повествовательных предложений характеризуется завершённой интонацией – ***F*** (Finality). Категория распознаётся по знакам: [.], [**…**], [*], [#], которые определяют её интонационный тип, обозначаемый при обработке текста, соответственно, символами:

– ***F0*** – интонация «точки» - [.],

– ***F1*** – интонация «многоточия» - […],

– ***F2*** – интонация «заголовка» - [*],

– ***F3*** – интонация «абзаца» - [#].

Кроме перечисленных выше основных пунктуационных типов интонации завершённости, реализующихся в последней синтагме предложения, внутри него могут присутствовать также два дополнительных пунктуационных типа интонации, характеризующихся различной степенью завершённости:

[1] Здесь и далее (в примерах 2-7) граница синтагмы в предложении обозначается символом «/»

086-Cirulnik_L:_Layout 1 18.05.2008 16:27 Page 565

– ***F4*** – интонация «точки с запятой» – [;],

– ***F5*** – интонация «вводности» – [ )], [,–], [–].

Интонация «вводности» реализуется при условии, что в предложении указанным знакам предшествовали, соответственно, знаки [( ], [,–], [–].

Внутри предложения могут присутствовать также 4 пунктуационных подтипа интонации, характеризующихся различной степенью незавершённости: ***N*** (Nonfinality):

– ***N0*** – интонация «запятой»- [,],

– ***N1*** – интонация «тире» - [–],

– ***N2*** – интонация «двоеточия» - [:],

– ***N3*** – интонация «предвводности»- [( ], [,–], [–].

Интонация «предводности» реализуется при условии, что за указанными знаками непосредственно следуют, соответственно, знаки [ )], [,–], [–].

В свою очередь пунктуационные синтагмы могут содержать лексические синтагмы с интонацией незавершённости следующих 3-х типов:

– ***N4*** – интонация «союза И»,

– ***N5*** – интонация «союза ИЛИ»,

– ***N6*** – интонация лексических синтагм – [ИС], [АБ], [РЧ], [НВ].

Далее, как само предложение, так и входящие в него пунктуационные и лексические синтагмы могут содержать неопределённое количество синтаксических синтагм [3] с характерной для них интонацией незавершенности:

– ***N7*** – интонация синтаксических синтагм.

(8) Примеры[2]:

Возможность победы ***БНФ*** *[N6] вызывает большие сомнения[F0].*

В пробирке оказалось 2 ***миллиона*** *[N6]* 350 ***тысяч*** *[N6] молекул белка[F0].*

Сегодня в 10 ***часов*** *[N6] 15* ***минут*** *[N6] 34* ***секунды****[F0].*

Он приехал в ***четверг*** *[N6] 20-го* ***июня*** *[N6] 7-го* ***года****[N6] навсегда[F0].*

Время от времени [N3] – для разрядки [F5] – он вставлял шутки[F0].

*Любой народ [N3]****, -*** *говорил он [F5]****, -*** *достоин уважения[F0].*

Описанные интонационные типы синтагм показаны на рис. 1.

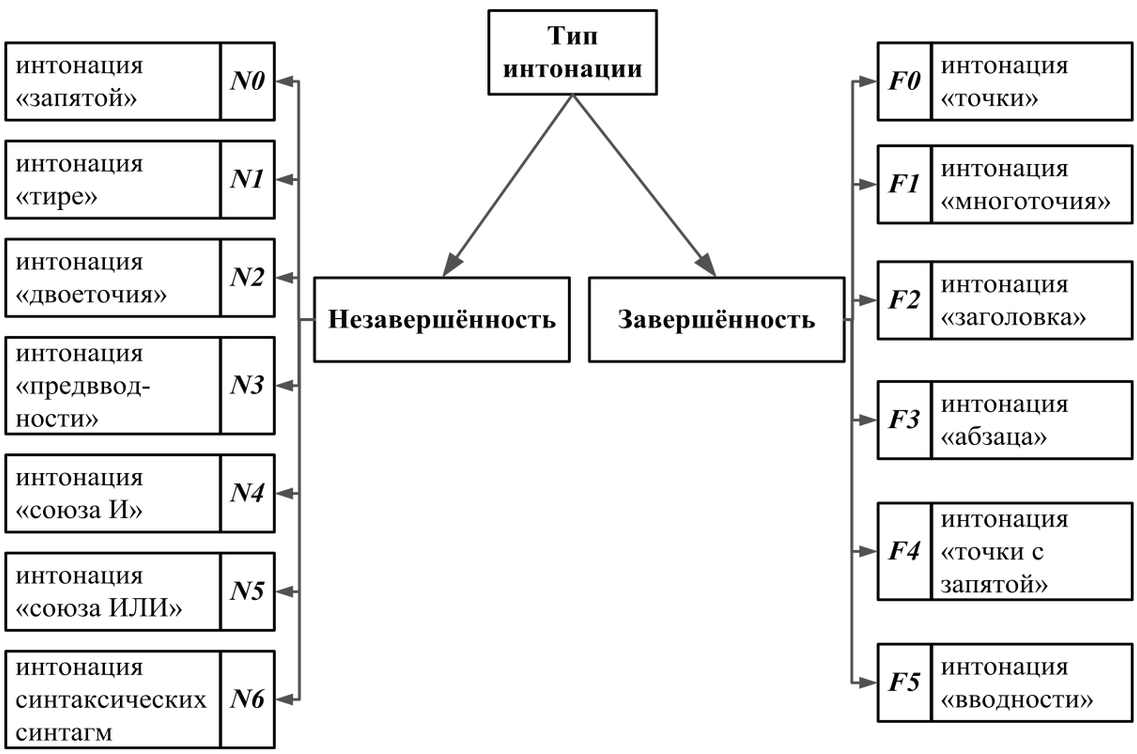


Рис.1. Интонационные типы синтагм повествовательных предложений

[2] Здесь и далее (в примерах 8-15) интонационный тип синтагмы указан в квадратных скобках.

086-Cirulnik_L:_Layout 1 18.05.2008 16:27 Page 566

3. Маркировка комбинаторных вариантов интонационного типа синтагм

Многие из рассмотренных выше интонационных типов пунктуационных и лексических синтагм могут иметь определённые комбинаторные варианты. Это замечание в наибольшей степени касается интонационных типов ***N0*** *[,]* и ***F0*** *[.]*. Причиной возникновения комбинаторных вариантов являются определённые различия в левом и правом контекстах анализируемой синтагмы, определяемые типом союзного слова, используемого наряду со знаком запятой для разделения синтагм. При этом комбинаторные варианты интонационного типа ***N0*** образуются за счёт различий в правом контексте ПС, а ***F0*** - за счёт различий в левом контексте. Запятой и союзом могут отделяться однородные члены внутри предложения, а также сложносочинённые и сложноподчинённые предложения. Рассмотрим подробнее особенности возникновения комбинаторных вариантов интонационных типов ***N0*** и ***F0***.

Можно выделить следующие основные варианты интонирования синтагм в зависимости от способа отделения однородных и обособленных членов предложения, а также сложносочиненных предложений друг от друга:

1. Однородные члены предложения, отделяемые запятой и следующими за ней соединительными или разделительными союзами: *и, ни...ни, или, либо, ли...ли, то...то,* и др.

Комбинаторный вариант (0) - ***N0.0***, ***F0.0.***

(9) Примеры:

И пращ [N0.0], и стрела [N0.0], и лукавый кинжал [F0.0].

За дождем не видно было ни моря [N0.0], ни неба [F0.0].

Гаврила либо сбежал [N0.0], либо утонул [F0.0].

Стало совсем темно [N0.0], и улица мало-помалу опустела [F0.0].

2. Однородные члены предложения, отделяемые запятой и следующими за ней противительными союзами: *а, но, да* (в значении «*но*»), *однако* и др.

Комбинаторный вариант (1) - ***N0.1***, ***F0.1.***

(10) Примеры:

На смелого собака лает[N0.1], а трусливого кусает [F0.1].

Он был силен [N0.1], да не умен [F0.1].

3. Обособленные члены предложения, отделяемые причастием.

Комбинаторный вариант (2) - ***N0.2***, ***F0.2.***

(11) Пример:

Внезапно он улетел [N0.2], встревоженный вихрем [F0.2].

4. Обособленные члены предложения, отделяемые деепричастием.

Комбинаторный вариант (3) - ***N0.3***, ***F0.3.***

(12) Пример:

Длинная стружка лезла из рубанка [N0.3], завиваясь штопором [F0.3].

5. Сложносочинённое предложение, отделяемое сочинительным союзом.

Комбинаторный вариант (4) - ***N0.4***, ***F0.4.***

(13) Пример:

Гости уехали *[N0.4]*, и в доме наступила тишина *[F0.4]*.

6. Сложноподчинённое предложение, отделяемое подчинительным союзом.

Комбинаторный вариант (5) - ***N0.5***, ***F0.5.***

(14) Пример:

Все заглядывали вперед [N0.5], где качалось красное знамя [F0.5].

Замечание. При отсутствии признаков, определяющих указанные выше интонационные варианты, второму индексу присваивается значение «6».

(15) Пример:

Впереди виднелись горы [N0.6], их вершины блестели [F0.6].

Предложенные правила маркировки комбинаторных вариантов представлены на рис. 2.

086-Cirulnik_L:_Layout 1 18.05.2008 16:27 Page 567

Рис.2. Правила маркировки комбинаторных вариантов интонационного типа синтагмы

4. Маркировка позиционных вариантов интонационного типа синтагм

В процессе синтеза речи особенно важно избежать так называемой «монотонности второго рода». Этот вид монотонности проявляется при использовании одних и тех же интонационных конструкций для двух или более идущих подряд синтагм одного интонационного типа. В естественной речи говорящий, как правило, стремится избежать такого рода монотонности путём варьирования интонационных параметров. Это замечание в наибольшей степени касается интонационных подтипов ***F0.0, N0.0***, для которых частота последовательного появления в текстах весьма значительна.

Определим минимально необходимый набор позиционных вариантов указанных интонационных типов.

Позиционные варианты интонации завершённости – ***F0.0***:

Позиционный вариант (0)

– ***F0.0.0***, при условии, что этот интонационный подтип встретился в абзаце впервые или в 3-й, 5-й и т.д. нечётный раз подряд, и так - вплоть до конца абзаца;

Позиционный вариант (1)

– ***F0.0.1***, при условии, что интонационный подтип ***F0.0*** встретился в абзаце во 2-й, 4-й, и т.д. чётный раз подряд, и так - вплоть до конца абзаца.

Позиционные варианты интонации незавершённости – ***N0.0***:

Позиционный вариант (0)

– ***N0.0.0*** при условии, что этот интонационный тип встретился в предложении впервые или в 3-й, 5-й и т.д. нечётный раз подряд раз подряд, и так - вплоть до конца предложения;

Позиционный вариант (1)

– ***N0.0.1***, при условии, что интонационный тип ***N1*** встретился в предложении во 2-й, 4-й, и т.д. чётный раз подряд, и так вплоть до конца предложения.

При необходимости подобным же образом возможно создание вариантов других интонационных типов, рассмотренных выше.

Предложенные правила маркировки позиционных вариантов представлены на рис. 3.

086-Cirulnik_L:_Layout 1 18.05.2008 16:27 Page 568

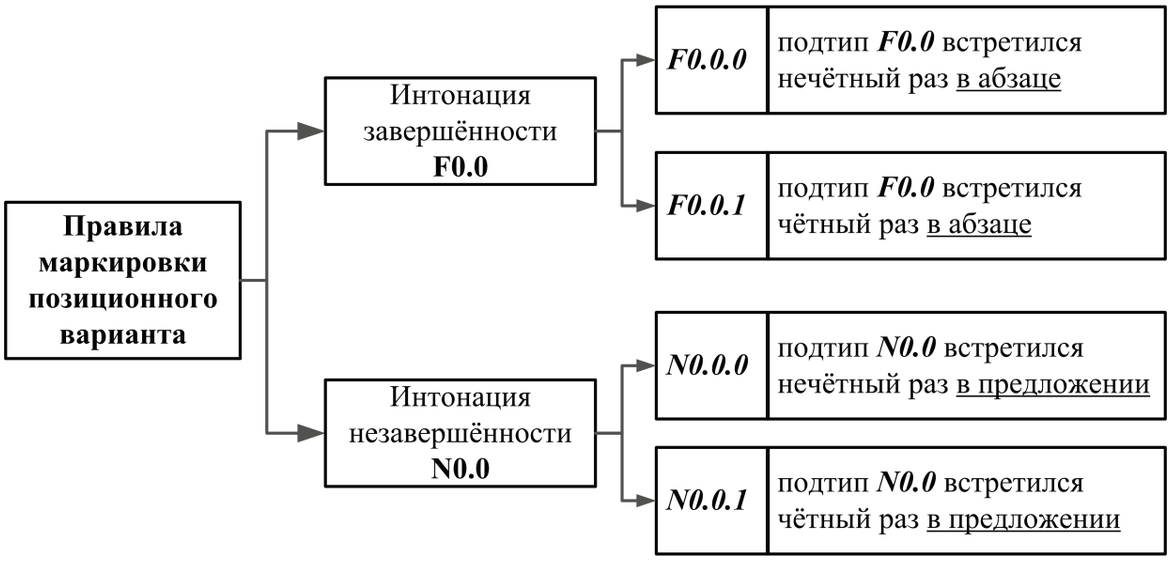


Рис.3. Правила маркировки позиционных вариантов интонационного типа синтагмы

В общем случае интонационные тип, подтип, комбинаторный и позиционный варианты каждой синтагмы в повествовательном предложении обозначаются буквой и следующими за ней тремя индексами *i*, *j*, *k* , которые означают следующее:

– буква – интонационный тип синтагмы: ***F*** (завершённый) или ***N*** (незавершённый);
– индекс *i* – интонационный подтип синтагмы: *0*, *1*, *2*, …
– индекс *j* – комбинаторный вариант подтипа синтагмы: *0*, *1*, *2*, …
– индекс *k* – позиционный вариант подтипа синтагмы: *0*, *1*, *2*, …

Заключение

Предложенный алгоритм позволяет сгенерировать достаточно большое количество вариантов синтагм завершённого и незавершённого типа, что обеспечивает устранение в синтезированной речи так называемой «монотонности второго рода».

Рассмотренный подход к маркировке интонационных подтипов синтагм в повествовательных предложениях, их комбинаторных и позиционных вариантов может быть использован также и при маркировке вопросительных и восклицательно-побудительных предложений.

Доклад будет проиллюстрирован образцами речи, синтезированной в соответствии с описанным алгоритмом.

Список литературы

1. Лобанов Б.М. и др. Алгоритмы синтеза просодических характеристик речи по тексту в системе ”Мультифон” // Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог’2007, М.: Издательский центр РГГУ, 2007. – С. 550-558.

2. Лобанов Б.М. Алгоритм сегментации текста на синтаксические синтагмы для синтеза речи // в наст. сб. трудов Диалог’2008.

087-Sharonov_I:_Layout 1 18.05.2008 16:27 Page 569

К ВОПРОСУ О РАЗГРАНИЧЕНИИ ЭМОЦИОНАЛЬНЫХ МЕЖДОМЕТИЙ И МОДАЛЬНЫХ ЧАСТИЦ

BORDERLINES BETWEEN EMOTIONAL INTERJECTIONS AND MODAL PARTICLES

Шаронов И.А. *(igor_sharonov@mail.ru)*
Российский государственный гуманитарный университет

Статья посвящена выявлению признаков, на основании которых эмоциональные междометия можно отграничить от различных разрядов частиц. К междометиям следует относить единицы в синтаксически изолированных позициях, являющиеся непроизвольными, неадресованными реактивными репликами, реакциями не только на языковые, но и на экстралингвистические стимулы.

Эмоциональные междометия и частицы относятся к периферии языковой системы. Они обладают набором общих и близких признаков. Перечислим некоторые из них.

1. Обе группы единиц не способны изменяться, они не располагают системой грамматических форм. Ср. первичные междометия *ах, ой, ба, ого* и частицы *уж, ли, же* и т. д. Некоторые единицы, например, *а, ага, о, ну, да* и др. в зависимости от контекста определяются то как междометия, то как частицы.
2. Междометия и частицы не обладают знаменательностью, не служат для называния предметов или признаков внешнего мира.
3. Значение междометий и некоторых частиц в очень большой степени зависит от интонации их произнесения [Кодзасов 1996; Кобозева 2006].
4. Некоторые (модальные) частицы выражают состояние субъекта, «позицию говорящего, то, как он относится к содержанию сообщаемого <…>, как квалифицирует события, о которых сообщает» [Шимчук, Щур 1999: 9]. Такие значения очень близки тем значениям, которые выражают эмоциональные междометия.
5. И междометия, и частицы имеют развернутые группы вторичных единиц, различить которые только по формальным показателям едва ли возможно. Ср. междометия: *вот это да; ничего себе; ишь ты (поди ж ты)* и частицы: *вот так-то; еще чего; то-то же; ну да* и т.п.

Все перечисленные признаки указывают на близость и слабую различимость двух групп единиц. В ряде западных лингвистических школ междометия определяются как подкласс частиц [Jespersen 1965, Schiffrin 1987, Fraser 1999 и др.].

В русистике междометие традиционно считается особой частью речи, стоящей вне деления на знаменательные и на служебные слова [Белошапкова и др. 1981: 256; Лекант и др. 1999: 194 и т. д.]. Однако в последнее время среди русских лингвистов раздаются голоса в пользу включения междометий в состав служебных слов, где, вероятно, они должны смыкаться с частицами [Ефремова 2001: 8].

Мы попробуем определить и обосновать границы между двумя рассматриваемыми классами, решая при этом не вопросы общей теории грамматики, разделения слов по частям речи, а выявляя специфические функции каждой из рассматриваемых групп слов.

Наш подход основан на представлении о междометиях как реактивных непроизвольных эмоциональных возгласах[1]. Опора на эти признаки позволяет делать сопоставительный анализ междометий и частиц на синтаксическом и семантико-прагматическом уровнях.

1. Синтаксический уровень

Чтобы утверждать, что единица относится к группе эмоциональных междометий, необходимо обнаружить ее в абсолютивной или в автономной синтаксической позиции. Ф. Амека жестко указывает на обязательность

[1] Междометия — это класс неизменяемых слов, служащих для нерасчлененного выражения чувств, ощущений, душевных, состояний и других (часто непроизвольных) эмоциональных и эмоционально-волевых реакций на окружающую действительность [РГ 1980, т. 1: 732].

087-Sharonov_I:_Layout 1 18.05.2008 16:27 Page 570

этого признака: «Междометия могут быть высказываниями сами по себе и всегда отделены паузой от высказываний, совместно с которыми они могут выступать. Они всегда имеют собственный интонационный контур» [Ameka 1992:108][2].

При таком подходе слова, внешне идентичные некоторым междометиям, однако используемые в неизолированных синтаксических позициях (*ай, ой, ах, ох, ух, эх*), должны относиться частицам, так как не имеют собственного интонационного контура и возможности передавать реакцию на внешний стимул.

В русистике данный подход представлен в [ГРЯ 1952, т. 1, § 1015; Розенталь 1999: 140 и др.]. Ему противостоит подход, в котором исследователи опираются на внешнюю форму перечисленных выше единиц [Шведова 1957, 1960; Германович 1966; РГ 1980; Середа 2005 и др.]. Опора на внешнее сходство единиц приводит к утверждению, что междометия могут функционировать не только в качестве эквивалента предложения, но также члена предложения и элемента конструкции.

В результате синтаксическая специфика междометий нивелируется, различие между ними и частицами, выступающими в предложении в качестве строевых элементов или модификаторов[3]., практически полностью стирается. Признак изолированности употребления позволяет отделить междометия от большинства частиц. Поэтому для общего грамматического описания и для решения прикладных задач первой подход представляется нам более продуктивным.

Однако полное размежевание рассматриваемых классов слов на основании синтаксического критерия оказывается невозможным. Среди частиц есть такие, которые также используются в синтаксически изолированных позициях, выступают как отдельные реплики. Поэтому для дальнейшего рассмотрения различий необходимо перейти на уровень содержательных характеристик рассматриваемых единиц.

2. Семантико-прагматический уровень

Среди репликовых частиц можно выделить две группы. Единицы первой используется как утвердительные и отрицательные реакции на вопрос, на предложение или на оценку собеседника. В репликах подтверждения, согласия используются частицы: *Да; Ага; Угу; Есть* (*Выполняйте. – Есть!*); *Точно; Так; Ну* и др. В репликах отрицания, несогласия, отказа используются частицы: *Нет; Нетушки; Никогда; Ни за что; Да ну; Ну прямо!; Ни боже мой; Черта с два*. Ср.:

Реплики подтверждения, согласия:

(1) *Т р и г о р и н. Должно быть, в этом озере много рыбы.*
*Н и н а. **Да** (А. Чехов. Чайка).*

(2) *«Благодаря тебе они перестали ходить лечиться». – «**Ага**, перестали!» – ухмыльнулся генерал (А. Чехов. Скука жизни).*

(3) *«Хорошая баба?» – «**Ну!**.. – Гринька весь засиял. – Бывает, примерзнешь где-нибудь в лесу – хоть волком вой. А как ее вспомнишь, так, может, не поверишь, сразу жарко становится» (В. Шукшин. Любавины).*

Реплики отрицания, несогласия, отказа:

(4) *П р. и н ц е с с а. Христиан-Теодор, прости меня... Останься или возьми меня с собой.*
*У ч е н ы й. **Нет**, принцесса (Е. Шварц. Тень).*

(5) *«Если ты сейчас же не появишься, мы будем считать, что ты сдался, проклятый дезертир». – «**Ни за что**, мессир!» – заорал кот и в ту же секунду вылез из-под кровати (М. Булгаков. Мастер и Маргарита).*

(6) *«А ты бы свою мебель попросил, – сказал Чохов. – Рояль». «**Черта с два**! – махнул рукой Воробейцев. – У наших попросишь!» (Э. Казакевич. Дом на площади).*

Как можно видеть, репликовые частицы играют в диалоге самостоятельную роль, имеют статус речевых действий: с их помощью говорящий *подтверждает* сказанное собеседником, *соглашается* с оценкой, *отказывает* собеседнику и т.д. Репликовые частицы обладают такими важными признаками высказывания, как интенциональность и адресованность. С семиотической точки зрения это сигналы, которые говорящий направляет конкретному адресату, сигналы, в которых закодирована определенная информация.

Междометные выкрики в силу спонтанности эмоциональной реакции не обладают признаками интенциональности и адресованности. Их использование не имеет жесткой привязки к диалогу. Междометия

[2] Interjections can be utterances by themselves and they are always separated by a pause from the other utterances with which they may co-occur. They always constitute an intonation unit by themselves.

[3] Роли и функционированию частиц в высказывании было уделено большое внимание в русистике последних десятилетий. См. работы [Николаева 1965; Широкова 1982; Rathmayr 1985; Баранов, Кобозева 1988; Кобозева 1990 и др.].

087-Sharonov_I:_Layout 1 18.05.2008 16:27 Page 571

часто произносятся в отсутствии собеседника или третьих лиц. Так, вовсе не требуют присутствия собеседника эмоциональные вскрики при сбое в деятельности (*ой, ах, черт, упс*), междометные реакции физиологического характера (*фу, тьфу, брр*), вздохи (*ах, ох, эх*) и т.п. Даже если человек использует их намеренно, с целью привлечь к себе внимание, например, начнет вздыхать или стенать, стоящий рядом «имеет право» не услышать такие выкрики, исходя из их ненаправленности на кого бы то ни было и, в частности, на него.

Основываясь на признаках неадресованности и непроизвольности, Ф. Амека относит междометия к ментальным, а не речевым актам. На основании этого признака исследователь разводит употребления *Yes* (*Да*) в устной речи на два разных класса. В зависимости от того, а) идет речь просто о поддакивании в процессе восприятия речи собеседника, или б) об утвердительном ответе на вопрос, подтверждении чего-либо, слово *Yes* интерпретируется исследователем принципиально по-разному. Ф. Амека пишет: «Как сигнал поддержки обратной связи в протекающем дискурсе *Yes* является ментальным актом (междометием – И. Ш.), а в качестве ответной реплики – это речевой акт» [Ameka 1992: 115][4].

В репликовых частицах часто можно обнаружить выражение эмоционального состояния говорящего. Однако оно отодвинуто «в тень», так как главным компонентом их значения является речевая интенция – намерение говорящего подтвердить или опровергнуть мнение собеседника, например, принять предложение или отказаться от него. Эмоциональный компонент рассматривается при анализе репликовых частиц как добавка к семантике речевого акта, не меняющая его значения (за исключением, может быть, иронических контекстов), а лишь модифицирующая его.

(7) – *Ты словно как будто и меня боишься...*
– ***Ну вот еще****! Зачем мне тебя бояться! Я вижу... понимаю... (А. Чехов. Беспокойный гость).*

(8) «*Ты бы хоть овец выгнала! – крикнула старуха. – Барыня!». «****Вот ещё!*** *Стану я на вас, иродов, работать», – проворчала Варвара, уходя в дом (А. Чехов. Бабы).*

(9) – *Переночуйте с Любой, а завтра утром напьётесь чаю и поедете.*
– *«****Вот ещё!****» – рассмеялась Наташа (В. Вересаев. Поветрия).*

Частица-реплика *Вот ещё!* является намеренным речевым актом, адресованным собеседнику, отвечает ожиданиям собеседника получить ответ, выражает возражение или отказ в просьбе. Синоним этой реплики – нейтральная частица *Нет*, от которой *Вот ещё* отличается экспрессивной характеристикой: пренебрежением, насмешкой или балагурством.

Междометия не предназначены для передачи речевых актов, хотя в некоторых случаях слушатель может вывести из них информацию, функционально равноценную речевому акту. Макс Мюллер писал: «одно краткое междометие может быть выразительнее, точнее, красноречивее длинной речи» [Мюллер 1865: 282]. Проиллюстрируем данное утверждение анекдотом:

(10) – *Что ты такой грустный?*
– *Понимаешь, я вчера попросил руку одной девушки, а она мне отказала.*
– *Ну, ты знаешь, если девушка говорит «нет», это еще ничего не…*
– *Она не сказала «нет». Она сказала: «За тебя? Тьфу!»*

Репликовые частицы второй группы почти неотличимы от эмоциональных междометий. Это реплики диалога, выражающие ментальные эмоциональные состояния говорящего. В академической Русской грамматике данные единицы описываются следующим образом: «В качестве реплик в диалоге функционируют и многие другие <…> частицы, выражающие непосредственную реакцию на слова собеседника» [РГ 1980, т.1: 729]. Авторы грамматики дают значения и приводят примеры употребления нескольких таких частиц. Ср.:

Сомнение:	– Лентяй он. – *Будто уж (Ну уж).*
Предупреждение:	– Я больше не буду. – *То-то.*
Удивление, осуждение:	– *Ну и ну!*

В приведенном списке обращает на себя внимание единица *Ну и ну*, которую чаще можно встретить в списках междометий. Ее включение в список репликовых частиц – демонстрация нечеткости границы между такими частицами и междометиями.

Рассмотрим еще единицы, которые относятся ко второй группе репликовых частиц.

НУ КАК ЖЕ

(11) *«А она была в каком чине?» – «Она была просто женщина без чинов». «****Ну как же****, – заволновалась Искрина еще больше. – Ведь он называет ее гением» (В. Войнович. Москва 2042).*

КАК ЭТО ТАК

[4] Ср. описание двух указанных употреблений Yes с позицией А. А. Шахматова, который после некоторых колебаний вывел реплику подтверждения Да из класса междометий» [Шахматов 1925: 194].

087-Sharonov_I:_Layout 1 18.05.2008 16:27 Page 572

(12) «*Сейчас заберите вещи: брюки, пальто, все, что вам нужно, – и вон из квартиры!». – «**Как это так?**» – искренне удивился Шариков. – «Вон из квартиры – сегодня», – монотонно повторил Филипп Филиппович, щурясь на свои ногти (М. Булгаков. Собачье сердце).*

При анализе языкового материала можно заметить, что такие «эмоциональные» частицы ограничены в употреблении по сравнению с эмоциональными междометиями. Данные частицы используются исключительно в диалоге. Трудно представить внедиалогическое употребление единиц *Будто уж; Ну уж; То-то,* а *Ну и ну* может употребляться как в диалоге, так и вне его как реакция на экстралингвистическую ситуацию. В первом из предлагаемых ниже примеров *Ну и ну* используется в диалоге, а во втором – как внедиалогическая реакция говорящего на объект восприятия.

(13) «*Да, забыл сообщить: праздник продолжается целую неделю!» – «**Ну и ну**, – позавидовали пассажиры» (Т. Креветко. До чего же славной бывает масленица! // Трамвай1990, № 3).*

(14) *Бодро открываю капот и направляю луч карманного фонарика на тыльную сторону фары. А как иначе – моторный отсек у «Фокуса» без подсветки. **Ну и ну**, теснотища! (В. Крючков. Вторая зима // За рулем 2004, № 4).*

Таким образом, оказывается возможным отграничить междометия от репликовых частиц второй группы на основании невозможности / возможности служить реакцией на неречевой, ситуативный стимул.

Если принять такое противопоставление как принципиальное, то близкие по смыслу единицы *Что такое* и *То есть как* должны будут разойтись по разным частям речи. Сопоставим возможности употребления этих единиц в диалоге и вне его. Поскольку единицы близки по значению, сопоставление можно провести на основе их замены в одном конкретном контексте.

В (15) даются употребления реплики *Что такое* в диалоге, где синонимические замены возможны, а в (16) – употребления этой реплики вне диалога, где такие замены затруднительны.

(15) – *Не дадут отцу-матери разговеться спокойно. Где Петька?*

– *Во все церкви пошел...*

– «***Что такое?** Ничего не понимаю. Вот я ему уши надеру, как вернется (Н. Тэффи. Семья разговляется).*

Здесь *Что такое* легко заменяется на *То есть как.*

(16) Употребления вне диалога

*Отыскав свой подъезд, он достал ключи. Ни один из них двери не отпер. «**Что такое...**» – сердито пробормотал он, и снова, стервенея, принялся совать*

(В. Набоков. Дар).

Здесь *Что такое* не может быть заменено на *То есть как.*

Итак, несмотря на смысловую близость, единицы *Что такое* и *То есть как* следует относить к разным частям речи – соответственно междометиям и частицам.

Предлагаемый способ разграничения междометий и репликовых частиц согласуется в большинстве случаев с результатами морфологической классификации в словаре Р.П. Рогожниковой [Рогожникова 1983], где единицы *Да ну; Вот именно; Ещё бы!; Ещё как* квалифицируются как частицы**,** а единицы *Ишь ты; Вот тебе (и) на; Вот тебе раз; Вот это да!* – как междометия. Вместе с тем единицы *Вот то-то* и *Да ну* во втором значении включены в данном словаре в состав междометий, хотя они используются исключительно в диалоге. Ср.:

ВОТ ТО-ТО, в знач. *междом.* (разг.). Подтверждает предыдущее высказывание.
– Может быть, я и ошибся, просто я его мало знаю. – Вот то-то!;
Вот то-то, все вы гордецы! (Грибоедов. Горе от ума).

ДА НУ (разг.). **II.** *междом.* Выражает удивление, изумление.
– Он сегодня пришёл без опоздания. – Да ну? Это на него не похоже.

Итак, при всей близости употребления репликовых эмоциональных частиц и эмоциональных междометий между двумя группами можно провести границу. Единицы, являющиеся намеренными и адресованными, использующимися исключительно в диалоге репликами должны рассматриваться отдельно от единиц, выражающих непроизвольные, неадресованные восклицания, служащие реакциями на любой внешний (необязательно речевой) стимул.

087-Sharonov_I:_Layout 1 18.05.2008 16:27 Page 573

Список литературы

1. Баранов А.Н., Кобозева И.М. Модальные частицы в ответах на вопрос // Прагматика и проблемы интенсиональности. М., 1988. С. 45–69.
2. Белошапкова В.А., Земская Е.А., Милославский И.Г., Панов М.В. Современный русский язык. М., 1981. 560 с.
3. Богданов С.И. Морфология неполнозначных слов в современном русском языке. СПб, 1997.
4. Германович А.И. Междометия русского языка. Пособие для учителя. Киев. 1966.
5. ГРЯ 1952 – Грамматика русского языка. АН СССР. М. Т. 1, 1952.
6. Ефремова Е.Ф. Толковый словарь служебных частей речи русского языка. М., 2001. С. 863.
7. Кобозева И.М. Прагмасемантическая аномальность высказывания и семантика модальных частиц // Логический анализ языка: Противоречивость и аномальность. М., 1990. С.194 – 203.
8. Кобозева И.М. Описание означающего дискурсивных слов в словаре: нереализованные возможности // Вестник Московского университета. Сер. 9. Филология. 2006. № 2. С. 37 – 56.
9. Кодзасов С.В. Семантико-фонетическое расщепление русских частиц и просодическая информация в словаре // Словарь. Грамматика. Текст. М., 1996.
10. Лекант П.А. Современный русский литературный язык. М., 1999. 462 с.
11. Мюллер М. Лекции по науке о языке. СПб., 1865.
12. Николаева Т.М. Функции частиц в высказывании: На материале славянских языков. М., 1965. 169 с.
13. РГ 1980 – Русская грамматика, в 2 т. М., 1980.
14. Рогожникова Р.П. Словарь сочетаний, эквивалентных слову. М., 1983. 144 с.
15. Розенталь Д.Э. Справочник по правописанию и литературной правке. М., 1999.
16. Середа Е.В. Морфология современного русского языка. Место междометий в системе частей речи. М., 2005. 159 с.
17. Шахматов А. А. Синтаксис русского языка. Вып. 1. М., 1925.
18. Шведова Н.Ю. Междометия как грамматически значимый элемент предложения в русской разговорной речи. ВЯ № 1. 1957. С.85–95.
19. Шведова Н.Ю. Очерки по синтаксису русской разговорной речи. М., 1960. 377 с.
20. Шимчук Э., Щур М. Словарь русских частиц. Peter Lang GmbH. Frankfurt am Main. 1999.
21. Широкова Е.Г. Частица И и некоторые функции усилительных частиц //Семантика служебных слов. Пермь, 1982. С.166–176.
22. Ameka F. Interjections: The universal yet neglected part of speech // Journal of Pragmatics. 1992. North-Holland. № 18. P. 101-118.
23. Fraser B. An approach to discourse markers // Journal of Pragmatics. 1990. № 14. P. 383-395.
24. Jespersen O. Language. London. 1965.
25. Schiffrin, D. Discourse Markers. Cambridge: Cambridge University Press. 1987.

088-Shemanaeva:_Layout 1 18.05.2008 16:28 Page 574

ГЛАГОЛЫ ПОГРУЖЕНИЯ: СЕМАНТИКА И СОЧЕТАЕМОСТЬ
VERBS OF GOING DOWN: SEMANTICS AND COMPATIBILITY

Шеманаева О.Ю. *(shemanaeva@yandex.ru)*
Российский государственный гуманитарный университет

Описываются глаголы вертикального движения вниз в некоторой среде, или глаголы погружения: увязнуть, тонуть, погрузиться, провалиться, ухнуть, уходить. Показаны параметры, релевантные для семантического описания этих глаголов, такие как контролируемость погружения, скорость, среда, субъект. Обсуждаются пути метафоризации – идея погружения в негативное состояние, идея избыточности и исчезновения.

1. Введение

Глаголы вертикального движения вниз в некоторой среде (*увязать / увязнуть, тонуть / утопать / утонуть, погружаться / погрузиться, проваливаться / провалиться, ухнуть, уходить / уйти*[1]) представляют из себя интересный квазисинонимический ряд, обладающий рядом общих свойств, как в пространственных, так и в метафорических употреблениях.

(1) *Босые ноги ее по щиколотку погрузились в теплый песок* (Л. Улицкая)

(2) *Посреди поляны, провалившись тонкими ногами в снег почти по колено, стоял и швырялся снежками мальчишка ростом с Игнатика* (В. Крапивин)

(3) *Ноги ушли по щиколотку в ил и по колено в воду* (В. Крапивин)

Остановимся на известных метафорических идеях, объединяющих эти глаголы. Во-первых, согласно известному принципу Дж. Лакоффа и М. Джонсона, BAD IS DOWN – ‘плохое – внизу’, т.е. все плохое, подчинение контролю, порочность, печаль и другие отрицательные эмоции ориентированы вниз [Лакофф, Джонсон 2004]. Глаголы погружения соответствуют этому принципу – речь идет о неконтролируемом движении вниз, метафорически – о погружении в плохие состояния, неприятные для человека события и плохие эмоции.

Действительно, из-за того, что действие неконтролируемо, оно оказывается нежелательным:

(4) *Я будто все время падал, и лишь в последний момент мне удавалось или одну ногу, или другую вытащить из этой бесконечной весенней грязи, я увязал в ней почти по щиколотку, но выдергивал, успевал подставить под тело и удерживал равновесие.* [Владимир Шаров. Воскрешение Лазаря (1997–2002)]

Эта идея поддерживается метафорическими употреблениями конструкции уровня – уровни *по горло* и *по уши* связаны с дискомфортными для человека ситуациями (*работы / хлопот по горло*) и с довольно длительными состояниями (*влюбился по уши*), из которых человеку самому не так просто выйти, ср. наблюдения Т.В.Булыгиной и А.Д.Шмелева над семантикой погружения в неприятные эмоции: «человек, погрузившийся в жидкость на известную глубину, оказывается лишен доступа к информации о внешнем мире и затруднен в своих движениях; и ему бывает нелегко сразу же вынырнуть на поверхность» [Булыгина, Шмелев 2000: 284].

Во-вторых, погружение (даже этимологически, ср. *груз*) связано с идеей тяжести, а «’тяжесть’ является одним из ключевых концептов внутренней сферы человека, поскольку это одно из главных обозначений ДИСКОМФОРТА» [Кустова 2004: 279-307]. Для некоторых глаголов погружения эта идея выступает на первый план при метафорических переносах.

Наконец, в-третьих, идея погружения связана с большим количеством вещества, окружающего погружающийся объект, и, следовательно, глаголы погружения участвуют в конструкциях со значением обилия, большого количества.

В данной работе мы рассмотрим эту группу, в том числе с точки зрения участия в конструкции уровня типа *мальчик*$_{Q}$ *провалился*$_{V}$ *в снег*$_{X}$ ***по колено***$_{Z}$ ***(Q V в X по Z,*** где **Q** – объект (прототипически – человек), **V** – предикат, **X** – среда, **Z** – часть тела человека, являющаяся верхней границей для среды), рассмотренной нами в [Шеманаева 2007] в рамках теории Грамматика конструкций [Fillmore et al. 1989] и работ Московской семантической школы. Глаголы движения вниз по известному метонимическому переходу от процесса

[1] Интересно, что глагол *углубиться*, хотя у него есть значение движения вниз, выражает чаще не погружение, а проникновение, ср. **углубился в воду vs. углубился в нору/ в чащу леса / в тайгу.* Семантически он связан в первую очередь не с «вертикальным», а с «горизонтальным» значением слова *глубокий*. Метафорический переход – ‘большая степень увлеченности чем-либо’ – у этого глагола сходен с переходами других предикатов погружения.

088-Shemanaeva:_Layout 1 18.05.2008 16:28 Page 575

к результату описывают не столько само движение, сколько его результат – нахождение в некотором слое (*снегу, воде, песке, грязи*) на определенной глубине (*по щиколотку, по колено, по пояс* и т.д.).

Для группы глаголов погружения важны такие параметры описания значения, как **скорость погружения, среда погружения, субъект ситуации**. В лексикографическом портрете каузативного глагола погружения ОКУНУТЬ И.Б.Левонтиной выделяются сходные параметры: **среда погружения** «жидкое или сыпучее вещество», продолжительность пребывания объекта в среде – речь идет о времени, как и в параметре скорость погружения – («сразу или через очень короткое время вынуть»), и особенность контакта объекта со средой («так, чтобы на объекте или в нем осталось какое-то количество этого вещества») [Левонтина 2000: 246].

Так как речь идет о среде и о погружающемся объекте (прототипически – человеке), то выделяется определенная часть объекта (уровень), до которой происходит его прототипически неконтролируемое погружение в среду. Уровень распространения среды определяется как степень глубины погружения объекта. Глаголы погружения в конструкции уровня фиксированы на результате движения вниз – т.е. на нахождении, а не на процессе движения.

Обратим внимание, что глаголы *увязнуть, ухнуть, уйти, утонуть, утопать* объединяет общая приставка *у-* . В значения этой приставки входит ‘направление движения **от** чего-либо’, ‘направление движения **внутрь** чего-либо’ и ‘**предел** действия’. См. сценарий глаголов погружения с приставкой *у-*, т.е. исходную и конечную ситуации и правила переходов между ними в [Горелик 2001][2], см. также исследование приставки *у-* [Зализняк Анна А. 2001].

2. Глаголы погружения[3]

2.1.*Увязнуть*

Увязнуть – результат погружения в грязь, ил, песок или другую вязкую, рыхлую, густую или мягкую, а не жидкую и не твердую, субстанцию, причем увязнуть можно как на достаточно маленьком уровне (*щиколотка*), так и на более глубоком – например, *увязнуть по грудь в болоте*. Субъектом глагола *увязнуть* может быть как сам человек, так и ноги человека, а также животные и средства передвижения.

(5) *Такой нехороший сон приснился. Такая башня узкая, и я стою у самой башни, почти вплотную, а ноги у меня по щиколотку в глине увязли.* (С. Юрский)[4]

(6) *Офицер хотел перескочить на берег прежде, нежели положили доску, и по колени увяз в грязи.* (А. Герцен, Записки одного молодого человека)

Глагол *увязнуть* подразумевает, что дальнейшее движение в этой среде затруднительно или невозможно, субъект (человек, его ноги, животные или средства передвижения) застревает в данной среде:

(7) *Телега по такой грязюке просто бы не прошла, увязла бы по ступицы.* [Евгений Лукин. Катали мы ваше солнце (1997)]

Если человек идет, *увязая* в некоторой среде, его движение является замедленным, затрудненным этим неконтролируемым медленным погружением.

Существенно, что субъект должен передвигаться: статичные объекты, не движущиеся сами в среде, а служащие ориентиром для движущейся среды, обычно не являются субъектами глагола увязать, ср. ?*дом увяз в грязи*.

В метафорическом значении этот глагол обозначает глубокое погружение в некоторую некомфортную ситуацию, в неприятное положение[5], и также может употребляться в конструкции уровня для метафорического усиления. Механизм метафорического переноса глаголов погружения обычно связан с метафоризацией среды:

(8) *Увяз я в пессимизме и решил отравиться.* (М. Горький, Пожары)

(9) *В этих подлостях и так уж / Я увяз по грудь.* (М. Цветаева. Царь-Девица)

(10) *Многие увязли в экзаменах и не до общения.* [Автогонки-3 // Форум forumsport.ru, 2005]

2.2.*Ухнуть*

Ухнуть – глагол, описывающий быстрое, моментальное неконтролируемое движение вниз (в словаре МАС

---

[2] В этой статье в данный ряд включается также глагол *угрязнуть*, ср. ***по колена*** *ж сам он в мать сыру землю угряз* (Гл.Успенский, цит. по МАС) [Горелик 2001: 51].

[3] В параллельном русско-венгерском словаре, посвященном моделям управления и глагольным конструкциям [Апресян, Палл 1982] из интересующих нас глаголов движения вниз приведены модели управления глаголов *войти / опустить / падать / погрузиться / провалиться / сойти / спустить / тонуть / утонуть / уйти.* В нашем исследовании мы рассмотрим семантические особенности глаголов погружения и метафор от них.

[4] Большинство примеров взято из Национального корпуса русского языка (www.ruscorpora.ru).

[5] Ср. пословицу *Коготок увяз – всей птичке пропасть.*

088-Shemanaeva:_Layout 1 18.05.2008 16:28 Page 576

помечен как просторечный, в НКРЯ встречается 413 раз в художественных и публицистических текстах). Субъектом глагола *ухнуть* может быть только субъект целиком, а не отдельные его части, например, только человек, а не ноги человека:

(11) *Осторожно, держась за кусты и выбирая место, куда встать, а то сразу ухнешь по колено в бурьян или частый крапивник.* (Ю. Домбровский)

Исходно этот глагол связан со звуком *ух*[6]: *ухать* – это 'производить этот звук', затем произошел метонимический переход к действию, сопровождающему звук движению, например, *эх, дубинушка, ухнем*; *ухнуть топором*, а потом и просто на некоторое движение (движение вниз), необязательно сопровождаемое звуком *ух*, но являющееся мгновенным (со скоростью звука *ух*), как правило, неконтролируемым и внезапным[7].

Отметим, что глагол в этом значении может быть как непереходным (*ухнул в пропасть*), так и переходным (*ухнул ребенка на пол*), причем в обоих случаях действие является неконтролируемым.

Это движение тоже связано с погружением, но необязательно в вещество – важна в первую очередь идея мгновенного неконтролируемого движения вниз, ср. *он ухнул с обрыва, он ухнул в пропасть*. Если движение контролируемо, употребление глагола *ухнуть* в этой ситуации затруднительно, ср.:

(12) *??Спортсмен ухнул с вышки в бассейн.*

(13) *??Мальчики поспорили, кто из них ухнет с вышки.*

У глагола *ухнуть*, в отличие от остальных глаголов, гораздо реже употребляется видовой коррелят, ср. *увязать – увязнуть*, *утопать – утонуть*, *погружаться – погрузиться*, *проваливаться – провалиться*, *уходить – уйти*, но *ухнуть – ??ухать* (в соответствии с результативной природой значения этого глагола)[8].

Для глагола *ухнуть* отсутствует метафорический переход погружения в неприятное состояние, свойственный другим глаголам погружения (*погрузиться / уйти / утонуть / увязнуть*) т.к. профилирующие идеи значения этого глагола – скорость и внезапность, а не длительность погружения и трудоемкость последующего движения наружу. Поэтому метафорический переход для этого глагола – не погружение в неприятную среду на длительное время, как у глаголов *погрузиться, увязнуть* и *уйти,* а внезапное исчезновение, ср. *их дружба ухнула* (пример из МАС)*, ухнуть все деньги,*

(14) *...все его три-четыре миллиона* ***ухнули****, и Полозов в 60 лет остался нищий.* [Н.Г. Чернышевский. Что делать? (1863)]

(15) *Перед отъездом несколько бодрых пробежек: по вопросу Галиной работы (работа* ***ухнула****, кто-то занял завидное место еще вчера).* [Александр Болдырев. Осадная запись (блокадный дневник) (1941-1948)]

2.3. *Провалиться*

Провалиться описывает переход от пребывания на относительно твердой поверхности к погружению в жидкую, вязкую или рыхлую субстанцию – ср. толкование в русско-венгерском словаре *провалиться* – 'прорвав своей тяжестью опору, быстро переместиться вниз' [Апресян, Палл 1982].

Так, на твердом снегу – насте – человек стоит, а в рыхлый снег проваливается (например, до определенного уровня – *по щиколотку / по колено / по пояс / по грудь / по уши*). Как у глагола *ухнуть*, так и у глагола *провалиться* субъектом является скорее человек, чем его ноги, ср. *я провалился по пояс в снег* vs. *?ноги проваливались по колено в снег*.

(16) *Климов сошел с дороги, пропуская машину, и тут же провалился по колено.* (В. Токарева. Кошка на дороге)

(17) *...Чижегов, чтобы не сорвалось, сошел в воду и, чертыхаясь, проваливаясь по колено и выше, повел вдоль берега* (Д. Гранин. Дождь в чужом городе)

В основе метафорического переноса лежит идея перехода от чего-то изначально известного (твердой поверхности в пространственном употреблении) во что-то неизвестное (новую среду), причем не постепенный, а мгновенный, ср. *он провалился в сон*. Другое переносное значение связано с переносом отрицательной оценки на некоторое абстрактное действие, при этом валентность среды поверхностно не заполняется, ср. *он провалился*

---

[6] Интересно сравнить значение глагола *ухнуть* с такими исходно звуковыми предикатными словами, *как бац, хлоп* и *трах*: в них заложена идея быстроты, неподготовленности, неожиданности действия [Шведова 2003: 265-267].

[7] Другие глаголы, исходно являющиеся звуковыми, развивают значение движения вниз: *загудеть / загреметь* – начинательные глаголы: *загудеть* = 'начать гудеть', *загреметь* = 'начать греметь', в переносном употреблении и тот, и другой глагол обозначают движение, не связанное со звуком, но мгновенное и неконтролируемое, оцениваемое как отрицательное. Ср. невозможные примеры **он загремел на новую высокооплачиваемую работу*. Ср. также *бухнуться, хлопнуться, плюхнуться, грохнуться* – глаголы мгновенного неконтролируемого движения вниз – падения, сопровождаемые соответствующим звуком.

[8] Процессные употребления глаголов *увязать, погружаться, тонуть, утопать, проваливаться*, допускают распространение наречием выше, ср.: *проваливаясь по колено и выше; тонул в снегу по пояс и выше.* Для ряда глаголов допустимо также наречие *глубже: проваливаясь по колено и глубже,* но не *дальше*, ср. *??проваливаясь по колено и дальше. Дальше* релевантно для глаголов направленного невертикального движения, например, *заходить в воду не дальше, чем по щиколотку.*

088-Shemanaeva:_Layout 1 18.05.2008 16:28 Page 577

на экзамене.

2.4. *Погрузиться*

Погрузиться описывает более медленное движение[9], чем *провалиться* и *ухнуть*. *Погрузиться* можно в жидкую, вязкую, рыхлую субстанцию: *воду, болото, песок*. Однако *снег* – нехарактерная среда для глагола *погрузиться*.

(18) *Когда на берегу прекратилась стрельба, он остановился, погрузившись по грудь в заросшее камышом болото.* [Василь Быков. Болото (2001)]

В отличие от остальных глаголов вертикального движения вниз, *погрузиться* может быть контролируемым:

(19) *Аркашка упивался, наслаждался, точно в июльскую жару погрузился по горло в прохладную воду и млел, и чуть шевелил пальцами ног.* [Василий Шукшин. Танцующий Шива]

(20) *Илья опустил голову и засмотрелся на лунную дорожку, пересекшую воды карьера от одного берега до другого, и ему внезапно, почти до физической боли в теле, захотелось погрузиться в эту прохладу по горло и поплыть.* [Дмитрий Липскеров. Последний сон разума (1999)]

Этот глагол допускает в качестве субъекта (как и глагол *увязнуть*) и человека, и ноги человека[10], причем для человека погружение будет более полное, ср.

Ноги погрузились в песок vs. *Я погрузился в священные воды древней реки*.

Впрочем, *погрузиться* – это не полное исчезновение человека с головой под водой, эту ситуацию обычно описывает глагол *окунуться*. Поэтому именно глагол *погрузиться*, а не *окунуться* употребляется в конструкции уровня, так как для него допустима градация погружения.

Плавность и возможная контролируемость действия у глагола *погрузиться* объясняют метафоры *город погрузился во тьму* и *мальчик погрузился в размышления, дом постепенно погружается в тишину* (пример из [Апресян, Палл, 1982]).

2.5. *Уйти*

Глагол *уйти* описывает не постепенное, а мгновенное неконтролируемое погружение самого человека, или одной его руки/ноги в некоторый другой объект. Этим объектом может быть как слой (*нога ушла по щиколотку в ил / землю / песок*), так и мягкий объемный объект, при этом направление движения, как у глагола углубиться, будет не ‘вниз’, а ‘вглубь’ (*ушел по плечи / головой в подушку*).

(21) *В воздухе пахнет первобытной гнилью, ноги по щиколотку уходят в рыхлую, прохладную землю.* [Фазиль Искандер. Дедушка]

(22) *Между тем Руська Доронин то и дело резко менял положение: он валился ничком, по самые плечи уходя в подушку[11], натягивая одеяло на голову и стаскивая с ног;*[А. Солженицын. В круге первом (т.1) (1968)]

В отличие от глаголов *идти* и *зайти,* которые употребляются в значении постепенного продвижения человека на глубину и – в результате – местонахождения (*зайти по колено в воду* = ‘начать находиться в воде, так что вода находится на уровне колен’), но в принципе сочетаются с движением в некоторым слое и без указания на уровень (*зайти в воду*), глагол *уйти* для субъекта ‘человек’ или ‘ноги человека’ приобретает значение мгновенного погружения **только** в конструкции уровня, ср. *ноги ушли по колено в снег / *ноги ушли в снег*[12]. Если выражение **по Z** не употребляется, то значение мгновенного погружения, которое мы хотели бы выразить (**я ушел в снег* как *я провалился в снег*) вступает в конфликт с контролируемым значением этого глагола, не направленным вниз (*я ушел в магазин*). Чтобы отменить это значение, надо обязательно вводить уточняющую конструкцию уровня: *Внезапно я ушел по колено в снег*. Отметим, что для нейтрализации основного значения лучше, когда уточняющая конструкция следует непосредственно за глаголом, ср.: *?я ушел в воду по колено* – конфликт с горизонтальным движением vs. *я ушел по колено в воду.* Для выражения смысла мгновенного погружения человека целиком, без указания на уровень, употребляется другой предлог: *ушел* ***под*** *воду,* или выражение *уйти с головой* (Ср. о неодушевленных субъектах *уйти под землю* – ‘полностью исчезнуть’).

Кроме одушевленных субъектов, у глагола *уйти* могут быть и неодушевленные субъекты, как статичные – здания (*сторожка ушла по окна в землю*), так и динамичные – средства передвижения и их части (*колеса по оси ушли в грязь*).

---

[9] *И вот, понимаете, страшный удар, борта затрещали, вода хлынула на палубу, и «Беда», рассеченная пополам, стала* ***медленно*** *погружаться в пучину.* [Александр Некрасов. Приключения капитана Врунгеля (1960-1980)].

[10] Вне конструкции уровня *погрузиться* обычно описывает движение вниз транспортных средств, управляемых человеком, как контролируемое (подводных лодок) так и неконтролируемое (кораблей).

[11] Ср. также метафорическое осмысление постельного белья как слоя в выражении *нырнуть под одеяло*.

[12] Заметим, однако, что *уйти в W* в значении погружения допустимо в метафорическом, а не в пространственном значении. Например, человек может *уйти в себя*, то есть глубоко погрузиться в свой внутренний мир (ср. [Булыгина, Шмелев 2000: 285]). (В отличие от таких пространственных примеров, как *уйти в магазин*, где идея погружения не рассматривается).

088-Shemanaeva:_Layout 1 18.05.2008 16:28 Page 578

Выше мы говорили про глагол *ухнуть*, что его профилирующая метафорическая идея – это исчезновение, связанное с мгновенным исчезновением, лежащим в основе пространственного значения. У глагола *уйти*, кроме погружения (*уйти в себя, уйти в работу*[13]), есть и сходный метафорический переход с идей исчезновения: *все ушло в песок* = ‘безвозвратно исчезло’.

2.6. *Тонуть / утопать*

Прототипическая среда глагола *тонуть* – вода, тонуть – это ‘погружаться в воду, идти на дно’ и ‘гибнуть, погружаясь в воду, на дно’ (МАС) (Подробнее о глаголах погружения в воду см. типологические исследования в проекте Aquamotion [Майсак, Рахилина 2007: 68-73]). Во втором значении среда меняется на нечто ‘вязкое, сыпучее, мягкое’, то есть имеется в виду погружение в песок, грязь, снег, даже темноту, но не в воду, так как в этом случае контекст вступает в конфликт с первым значением.

Тонуть в воде можно, только если человек полностью находится в воде и не может стоять, а если он стоит в воде по колено, про него нельзя сказать, что он тонет в воде по колено, даже если эта вода находится не в естественном водоеме, а в закрытом помещении – в лодке, в комнате и т.д. Поэтому у глагола *тонуть* во втором значении – состояния погружения в некоторый слой, кроме ограничений на уровень, есть ограничения и на слой: им не может быть вода, так как *тонуть* в воде понимается как ‘гибнуть’.

(23) *Снег был свеж и настолько глубок, что я тонул по колено.* (И.А. Бунин. Воспоминания)

(24) *Они тонули в пыли — теплой серой или горячей черной — по щиколку, наслажденьем было медленно брести, взрывая тут же опадающие крохотные воронки—бурунчики.* [Александр Чудаков. Ложится мгла на старые ступени // «Знамя», № 10–11, 2000]

Глагол *утопать* устроен почти так же, как глагол *тонуть*, только буквальное, первое значение – погружение в воду скорее является устаревшим, и *утопать* чаще используется как погружение в некоторый слой и в значении ‘иметь что-либо в избытке’, например, *утопать в слезах* или *утопать в зелени*[1].

(25) *Москвичи, передвигавшиеся по улицам, утопая в жидкой грязи по щиколотку, получили возможность поднимать глаза к небу и услаждать взор изысканными строениями* (Московский комсомолец’97)

(26) *Какой-нибудь сияющий огнями лайнер из Одессы, как призрак, бесшумно проплывал ночью в сторону Севастополя и Ялты, с него не доносилась до наших домишек, утопавших по крыши в подсолнухах и кукурузе, даже музыка* (М. Панин)

И *тонуть*, и *утопать* не являются глаголами мгновенного действия. Изначально они описывают некоторый процесс, но в конструкции уровня, как мы уже говорили, важен только результат этого процесса, состояние объекта. Так, *он тонет* – это предельный процесс, а *он тонет по колено в пыли* – это состояние. Таким образом, в этом значении из класса событий глагол *тонуть* переходит в класс состояний.

(27) *Пришел ко мне с кагалой: Сделай, батя, нам лыжи, в школу опаздываем, по шею в снегу тонем.* (В. Солоухин)

(28) *А грязь, вернее жидкая глина, была по щиколотку, не улицы, а глиняные реки, строящийся завод и город утопали в них.* [Григорий Бакланов. Мой генерал // «Знамя» №9, 1999]

3. Заключение

В группе глаголов погружения мы показали, какие функциональные признаки важны для классификации этих глаголов и могут объяснять различия в их метафоризации помимо трех общих главных концептов: 1) ВНИЗ – это плохо (дискомфорт, тяжесть); 2) движение ВНИЗ – это исчезновение; 3) ПОГРУЖЕНИЕ во что-либо – это избыток (как плохого, так и хорошего).

Релевантные признаки, выделенные в работе для описания глаголов погружения, представлены в таблице 1.

[13] Я зачитался. Я читал давно. / С тех пор, как дождь пошел хлестать в окно. / Весь с головою в чтение уйдя, / Не слышал я дождя. (Б.Пастернак)

[14] Часто метафорическое утопать – это иметь в избытке что-либо хорошее: утопать в роскоши, утопать в удовольствиях. Для негативного отношения используется глагол погрязнуть с прозрачной внутренней структурой, соотносящейся со словом грязь: погрязнуть в пороке / в развлечениях / в грехах.

088-Shemanaeva:_Layout 1 18.05.2008 16:28 Page 579

	Скорость погружения	**Среда погружения**	**Субъект погружения**	**Контролируемость**
увязнуть	нейтральная	вязкая, рыхлая (*снег, *вода)	ноги, человек, животные, средства передвижения	-
ухнуть	большая	вязкая, рыхлая, жидкая	человек	-
провалиться	большая	вязкая, рыхлая, жидкая	ноги, человек, животные, средства передвижения	-
погрузиться	маленькая	вязкая, рыхлая, жидкая *(*снег*)	ноги, человек, животные, средства передвижения	- / +
уйти	большая (движущиеся субъекты) / нейтральная (статичные субъекты)	вязкая, рыхлая, жидкая	ноги, человек, животные, средства передвижения, здания	-
тонуть	нейтральная	вязкая, рыхлая, жидкая (*вода)	ноги, человек, здания	-
утопать	нейтральная	вязкая, рыхлая, жидкая (*вода)	ноги, человек, здания	-

Таблица.1. Релевантные признаки фрейма для глаголов погружения

Некоторые релевантные метафорические сдвиги представлены в таблице 2.

	Метафора нахождения в негативном состоянии	Метафора избыточности	Метафора исчезновения
увязнуть	+	+	-
ухнуть	-	-	+
провалиться	-	-	+
погрузиться	+	+	-
уйти	-	-	+
тонуть	+	+	-
утопать	-	+	-

Таблица.2. Метафорические сдвиги глаголов погружения

Список литературы

1. Апресян Ю.Д., Палл Э. Русский глагол – венгерский глагол. Управление и сочетаемость. Будапешт, 1982.

2. Булыгина Т.В., Шмелев А.Д. Перемещение в пространстве как метафора эмоций. // Языки пространств. Логический анализ языка. Под ред. Н.Д.Арутюновой, И.Б.Левонтиной. М.: «Языки русской культуры», 2000. С. 277-288.

3. Горелик Е.В. Описание глагольной приставки у- // Московский лингвистический журнал. М.: РГГУ, №5/1, 2001. – С. 37-68.

4. Зализняк Анна А. Семантическая деривация в значении приставки У- // Московский лингвистический журнал, 5/1, 2001. – С. 69-85.

088-Shemanaeva:_Layout 1 18.05.2008 16:28 Page 580

5. Кустова Г.И.Типы производных значений и механизмы языкового расширения. М.: Языки славянской культуры, 2004. С.279-307.

6. Лакофф Дж., Джонсон М. Метафоры, которыми мы живем: пер. с англ. / Под ред. и с предисл. А.Н.Баранова. – М.:Едиториал УРСС, 2004.

7. Левонтина И.Б. Словарная статья ОКУНУТЬ, МАКНУТЬ, ОБМАКНУТЬ. // НОСС, 2000. С.246-248.

8. Майсак Т.А., Рахилина Е.В. Глаголы движения и нахождения в воде: лексические системы и семантические параметры // Глаголы движения в воде: лексическая типология. Под ред. Т.А.Майсака, Е.В.Рахилиной. М.: «Индрик», 2007. С.27-76.

9. МАС – Малый академический словарь русского языка в четырех томах под ред. А.П.Евгеньевой. (90 тыс.слов). М.: Русский язык, 1981.

10. Шеманаева О.Ю. Конструкции измерения уровня с предлогом по в русском языке. // М.: НТИ, 2007. Серия 2, №4, С.35-45.

11.Fillmore Ch., Kay P. & O'Connor M. Regularity and idiomaticity in grammatical construcions: the case of let alone. Language 63(3): 501-38. 1988.

089-Shmelevy_AE:_Layout 1 18.05.2008 16:29 Page 581

«МЫ» ИЛИ «ДРУГИЕ»: ИМИТАЦИЯ УКРАИНСКОЙ РЕЧИ В РУССКОМ АНЕКДОТЕ*

"WE" AND "OTHERS": THE SIMULATION OF UKRAINIAN SPEECH IN RUSSIAN JOKES

Шмелева Е.Я. *(eshkind@mail.ru),* ***Шмелев А.Д****. (shmelev.alexei@gmail.com)*
Институт русского языка им. В.В. Виноградова

Имитация украинской речи, насмешки над украинским языком и украинскими именами очевидным образом свидетельствуют, что анекдоты об украинцах рождаются в русском языковом окружении. В статье делается попытка выявить связи между сюжетом анекдота, «языковыми масками» его персонажей и этностереотипами.

Вступительные замечания

Как известно, жанр «этнического» анекдота присутствует в фольклоре самых разных народов. Поскольку культурные стереотипы разных народов обычно не совпадают, специфические черты поведения народов-соседей (в анекдотах, как правило, к тому же преувеличенные) часто воспринимаются как смешные, глупые, не соответствующие собственной, «правильной» норме поведения. Особенно смешными кажутся нарушения речевых стереотипов родного языка рассказчика и слушателей анекдота. Как это ни парадоксально, героями самых злых «этнических» анекдотов обычно становятся представители народов, весьма близких рассказчику по культуре, истории и языку. Испанцы рассказывают анекдоты о португальцах, французы о бельгийцах-франкофонах, а голландцы – о бельгийцах, говорящих на фламандском языке. Так и истории о малороссах бытовали в русском фольклоре еще до того, как в русской речевой культуре сложился жанр современного анекдота. Украинцы (*хохлы*) были, наряду с *грузинами*, *чукчами* и *евреями*, персонажами советских «этнических» анекдотов.

Не прервалась эта традиция и в постсоветское время. И после того как Украина стала независимым государством, в России продолжают рассказывать анекдоты об украинцах, причем в этих анекдотах практически не появилось ни новых мотивов, ни сюжетов. Некоторым исключением являются анекдоты, в которых в том или ином виде выражается сомнение в жизнеспособности Украины как независимого («самостийного») государства и украинского языка как государственного.

Конечно, в Интернете постоянно появляются новые политические анекдоты, в которых речь идет об украинских политиках, внутренней и внешней политике Украины, о российско-украинских отношениях, но эти анекдоты не имеют массового распространения и не вносят ничего нового в образ *хохла* – постоянного персонажа русских анекдотов на протяжении уже более ста лет[1].

Речевые характеристики и этнические стереотипы

Как и прочие персонажи этнических анекдотов, украинцы обладают легко опознаваемой речевой маской. Рассказчику нет необходимости сообщать, что в анекдоте, который он рассказывает, речь идет об украинцах, – достаточно имитировать речь украинцев как героев анекдотов. Поэтому ключевым для понимания анекдотов об украинцах (как и других этнических анекдотов) оказывается описание речевых масок их персонажей.

В этом отношении персонажи этнических анекдотов отличаются от «иностранцев». *Американцы*, *французы*, *немцы*, *англичане*, *японцы* не обладают легко опознаваемой речевой маской в качестве персонажей анекдота, хотя их речь может характеризоваться отдельными специфическими сигналами (междометие *оля-ля* в речи *француза*, обращение *сэр* в речи *англичанина*, использование свистящих вместо шипящих в речи *японца*).

* Исследование выполнено при финансовой поддержке РГНФ в рамках научно-исследовательского проекта РГНФ («Образ России и русских по языковым данным»), проект № 06-04-00591а

[1] Стоит сделать оговорку, которая может показаться излишней. Часто, говоря об этнических анекдотах, рассматривают их персонажей как объект насмешки и характеризуют эти анекдоты как свидетельство ксенофобии тех, кто их рассказывает (ср., в частности, книгу [Draitser 1998]).Такой подход излишне прямолинеен и вносит неуместный оценочный подход, препятствующий объективному анализу материала.

089-Shmelevy_AE:_Layout 1 18.05.2008 16:29 Page 582

Отметим, что в легко опознаваемой речевой маске «иностранцев» как героев анекдотов нет необходимости. «Иностранцы», как правило, выступают в роли персонажей мультинациональных анекдотов и называются посредством соответствующего этнонима (исключения – анекдоты из «французской» или «английской» жизни – относительно немногочисленны, и примечательно, что как раз в них специфические сигналы, характеризующие речь персонажа, становятся почти обязательны). Последним в мультинациональных анекдотах обычно действует (или говорит) представитель родного народа рассказчика (соответственно, в русских анекдотах – *русский*), именно он действует самым необычным, неожиданным образом, его слова или действия составляют то, что называется «солью» анекдота[2].

Обычно персонажи «этнических» анекдотов: *грузины*, *чукчи*, *евреи* или *эстонцы* – говорят по-русски, а их этноязыковая принадлежность подчеркивается имитацией акцента и использованием характерных частиц (*да? однако*, *таки*). При этом любопытно, что особенности речи этих персонажей оттеняют специфические черты национального характера. Так, *грузин*у свойственна установка на контакт и взаимопонимание с собеседником: грузин в анекдотах обращается на «ты» даже к незнакомому собеседнику; в его речи преобладают побудительные предложения и риторические вопросы; в конце каждого предложения он добавляет вопросительное «*да?*» – даже в тех случаях, когда это противоречит русской литературной норме. Непременное *однако* в речи *чукчи* отвечает общему внутреннему состоянию этого героя анекдотов – его пассивному изумлению перед окружающим миром (недоумение, вызванное встречей с незнакомой цивилизацией, подчеркивается и тем, что чукча всегда говорит о себе в третьем лице: он как бы глядит на себя со стороны и удивляется: как это он оказался в таком странном мире)[3].

Восприятие украинской речи как «неправильной» русской речи

На этом фоне речь украинцев – героев русских «этнических» анекдотов – отличают весьма своеобразные языковые характеристики: *украинец* в анекдоте чаще всего говорит не на «русском как иностранном», а «по-украински» – в той мере, в какой рассказчик в состоянии имитировать украинскую речь. Иными словами, речь *грузина*, *чукчи*, *еврея* или *эстонца* из анекдота – это *русская* речь с теми или иными отклонениями от нормы, обусловленными сознательной установкой рассказчика на имитацию русской речи человека, говорящего на неродном языке. Напротив того, речь «анекдотического» *украинца* – это *украинская* речь с отклонениями, которые обусловлены не сознательной установкой рассказчика, а тем, что рассказчик, как правило, не в достаточной мере владеет украинским языком[4].

По-видимому, русскоговорящему рассказчику украинский язык кажется «неправильным» русским. Это ощущение поддерживается тем, что в русских анекдотах чаще всего имитируются черты украинской речи (*мовы*), наиболее характерные для южнорусских диалектов. Сюда относится произнесение фрикативного *г* на месте русского литературного взрывного *г*, *и* на месте русского литературного *е* и на месте русского литературного *о* в новых закрытых слогах, использование союза *що* вместо русского литературного *что*. По-видимому, русское массовое сознание воспринимает все это не как особенности чужого самостоятельного языка, а как характеристики русской диалектной (т. е. «нелитературной») речи, иными словами – речи необразованных людей, не овладевших произносительными нормами русского литературного языка.

То, что украинцы говорят на языке, близкородственном русскому, но каком-то «неправильном», в котором многие слова, хотя звучат похоже, имеют другое значение, иногда составляет самую «соль» анекдота, В русскоязычной среде такие анекдоты рассказчику часто приходится комментировать:

(1) «На всемирном конгрессе женщин дамы обсуждали, как наказать мужчин за невнимательность к ним. Решили, что надо ограничить ласки. Долго спорили, до какой степени можно ограничить, и в конце концов постановили, что будут дарить им ласки только три раза в неделю. Приняли единогласно. «Вопросы, предложения есть?» Встает хохлушка: «У мэнэ есть спытання до конгресу. Що три раза у нэдилю, то я

[2] Разумеется, речь идет о «стандартном» случае рассказывания анекдота. При пересказе «чужого» анекдота (напр., финского анекдота по-русски, или русского анекдота по-шведски) последним действует или говорит представитель того народа, которому этот анекдот «принадлежит».

[3] Подробнее о связи речевых характеристик персонажей анекдотов: *грузин, чукчей, евреев, эстонцев* – и соответствующих этностереотипов см. в ряде наших предшествующих публикаций – напр., в уже упомянутой книге [Шмелева, Шмелев 2002: 47-63]. Заметим, что с этой точки зрения *новые русские* сходны по своим характеристикам с персонажами этнических анекдотов.

[4] Поскольку речь *украинцев* в русских анекдотах – это не подлинная украинская речь, а лишь ее имитация, не отвечающая нормам украинского языка, при ее передаче в рамках данной статьи мы, как правило, ограничиваемся средствами русской графики.

[5] К сходным выводам приходит Е. Е. Левкиевская в заметке «Украинцы глазами русских: эволюция этнокультурных и языковых стереотипов», опубликованной на сайте МИОН (http://www.iriss.ru/attach_download?object_id=000150070125&attach_id=000264). Ср. также [Draitser 1998: 67-70].

089-Shmelevy_AE:_Layout 1 18.05.2008 16:29 Page 583

согласная. А як будэ у будни?»

Обычно после этого анекдота рассказчик поясняет, что по-украински *неділя* – это воскресенье. Заметим еще, что прилагательного *согласный*, насколько мы знаем, в украинском языке нет, но русского рассказчика это нисколько не смущает: для имитации украинской речи достаточно, чтобы прозвучало несколько слов, которые играют роль своего рода условных сигналов.

Характер украинца как героя анекдота

При этом особенности речи *украинцев*, как и всех прочих персонажей этнических анекдотов, подчеркивают особенности их характера и поведения. Диалектная речь – это деревенская речь; соответственно, диалектные черты в речи украинца в анекдотах подчеркивают, что это типичный деревенский житель, необразованный, прижимистый и неопрятный, но не лишенный своеобразной хитрости и силы. Во многих отношениях стереотип украинца в русских анекдотах повторяет автостереотип русских, представленный в преувеличенном виде.

Украинский мужчина (*чоловік*) превыше всего ставит чинную мужскую беседу за бутылкой. Занятие любовью рассматривается им как несерьезное, недостойное настоящего мужчины. Но при этом всем в доме заправляет, безусловно, жена (как, впрочем, и у русских, насколько можно судить по русским семейным анекдотам):

(2) Женился татарин на хохлушке. Говорит ей: «Я когда приду домой и у меня тюбетейка на левый бок сдвинута, я тебя любить буду, целовать буду, деньги давать буду. А если у меня тюбетейка на правый бок сдвинута, значит злой-злой буду». А хохлушка ему отвечает: «Если ты пришел домой и видишь, что у меня руки на груди сложены, я тебя любить буду, целовать буду, борщом кормить буду, а если ты пришел и видишь, что у меня руки в боки – мне все равно, на какой бок у тебя тюбетейка сдвинута!

Замечательно, что рассказывая этот анекдот, обычно не пытаются имитировать украинскую речь: здесь русский язык выступает как «средство межнационального общения».

Украинцы и советская власть

Украинцы, как и другие персонажи советских этнических анекдотов, не любят советскую власть и коммунистическую партию. Но в отличие от всех остальных, они способны даже на реальное сопротивление:

(3) Во время войны перед боем солдат-украинец пишет записку: «Якщо вбьють, прошу вважити мене комунистом, а якщо ни, то ни».

(4) Встретились два украинца. Один рассказывает другому: «Був я у Испании на корыди, та бачив, як одного дурня бык на рога пидняв». – «А за що вин того?» – «А тоби перед рогами коммуняцким стягом махаты?»

(5) Дед Панас и внук пьют чай. Внук только что вернулся из сельпо, и расспрашивает деда. «Диду, а правду люды кажут, що у тэбэ в лису три танка е?» – «Брэшуть, внучку, брэшуть». – «Диду, а правду люды кажут, що у тэбэ в лису пять пушек?» – «Брэшуть, внучку, брэшуть». – «А шо у тим озэрку пидводна лодка плавае?» – «От чого нэма, того нэма…»

При этом во многих анекдотах нелюбовь к советской власти для украинцев оказывается неотделимой от нелюбви к тем, кто эту власть воплощал, – к русским (*москалям*). Само желание говорить на своем «ненастоящем», «искаженном русском» языке в анекдотах часто объясняется злонамеренностью украинцев и их нелюбовью к *москалям*.

Украинцы и «москали»

Этот стереотип вызвал к жизни особый жанр – псевдоукраинские анекдоты о нелюбви к *москалям* и к русскому языку (при этом *москали* в анекдотах такого рода могут уже никак не ассоциироваться с советской властью), например:

(6) «Петро, москали в космос полетили!» – «Що, уси?»

(7) Приехал москаль на Западенщину, нужно ему с вокзала до аэропорта на автобусе проехаться. Подходит он к дедку и спрашивает: «Дед, а где здись останивка?» – «Не останивка, а зупынка, а ты, москалику, вже прыйихав…»

(8) Бандеровец на могиле сына: «Чи я тэбэ не народыв, чи я тэбэ не годував, чи я тэбэ до Универсытэту не видпустыв, чи я тоби гроши не видсылав!? А ты прыйихав и що ты мени сказав? «Здравствуйте, папа!»»

089-Shmelevy_AE:_Layout 1 18.05.2008 16:29 Page 584

Многие из этих анекдотов родились в русской речевой среде (хотя, возможно, в среде русских, живущих на Украине), и это видно не только из того, что украинская речь зачастую имитируется в них довольно неумело (а анекдоты, требующие определенного уровня владения украинским языком, в русской среде чаще всего пересказывают «в сокращении»). Недаром постоянная их тема – насмешки украинцев над русским языком, русскими фамилиями. Ведь смешным это кажется именно носителям русского языка: мол, украинцы сами говорят на таком потешном языке, у них ужасно несуразные фамилии, а они еще смеются над нашим «великим и могучим» русским языком:

(9) Встречаются два хохла: Свербыгузенько и Черезплетеньходько. Один достает газету: «Бачь, яка чудна фамилия». – «Яка, яка?» – «Гу-у-усев».

(10) «Петро, чуешь, як москали наше пыво кличуть?» – «Як?» – «Пи-и-во!» – «Повбивав бы!»

Другие народы в «анекдотах об украинцах»

Помимо *москалей*, украинцы – герои русских анекдотов – не любят также и евреев (*жидов*). Любопытно, что помимо *москалей* и *жидов*, во многих анекдотах о *хохлах* фигурирует еще один довольно неожиданный персонаж – *негр*:

(11) Едет негр в Киеве в метро. Читает газету «Молодь Украины» на украинском языке. Рядом стоит здоровый хохол, спрашивает негра. «Може, ты хочешь сказаты, що ты розумиешь, що тут напысано?» – «Так, я розумию». – «Може, ты хочешь сказаты, що ты живешь у Кыиви?» – «Так, я живу у Кыиви». – «Може, ты хочешь сказаты, що ты украинец?» – «Так, я украинец». – «А хто ж я тоди такый?!» – «А бис тэбэ знае, – отвечает негр, – або жид, або москаль».

(12) Едет в поезде (в одном купе) негр и хохол. Едут-едут, пришло время ужина. У негра с собой ни фига нет. Достает хохол сало, чеснок, лук, черный хлеб и т. д., начинает есть. Негр на него смотрит, смотрит, уже слюни текут. Тут хохол к нему поворачивается и говорит: «Извиняйте, бананив немае».

(13) Заходит негр с обезьяной в автобус. Она от него – шасть и убежала… Негр: «Микки, Микки…» Хохол поймал обезьяну и говорит ей: «Мыкола, чому ж ты батьку-то не слухаешь?»

По-видимому, украинец, как типичная «деревенщина», негров (как и обезьян) никогда не видел, поэтому считает, что негры питаются только бананами, не умеют читать и т. п.[6]

Заключительные замечания

Выше мы отмечали, что в мультинациональных анекдотах последним обычно действует или говорит представитель родного народа рассказчика, слова или действия которого и составляют то, что называется «солью» анекдота. Любопытно, что в ряде русских мультинациональных анекдотов, в которых фигурируют украинцы, именно они оказываются последними в ряду персонажей, т. е. занимают место, предназначенное для *русских* (а, скажем, анекдотов, в которых это место занимают *грузины*, *чукчи* или *эстонцы*, довольно мало).

Может показаться, что в указанном отношении украинцы схожи с *евреями*, которые также могут выступать последними в русских мультинациональных анекдотах. Однако представляется, что источник такого «отклонения от нормы» для *евреев* и украинцев различен.

Относительно большинства анекдотов, в которых место, предназначенное для представителя родного народа рассказчика, занимают *евреи*, можно предположить, что они родились или активно рассказывались в еврейской среде (во многих случаях это достоверно известно). В отношении украинцев причина скорее в том, что украинцы рассматриваются в русских анекдотах как своего рода «русские». Э. Драйцер в книге о русских этнических анекдотах писал, что украинцам и чукчам в этих анекдотах приписываются те же черты, которые русские приписывают самим себе [Draitser 1998: 63-74, 94-100]; один из разделов главы, посвященной русским анекдотам о чукчах даже назывался «Чукча как русский» [*The Chukchi as a Russian*]. Пожалуй, в отношении *чукчей* этот вывод не подтверждается никакими языковыми данными; напротив того, имитация речи украинцев в анекдотах изображает их именно как русских, хотя и «неправильных».

Список литературы:

1. Шмелева Е.Я., Шмелев А.Д. Русский анекдот. Текст и речевой жанр. // М.: Языки славянской культуры, 2002.
2. Draitser E. Taking penguins to the movies: ethic humor in Russia. Detroit: Wayne State University Press, 1998.

[3] Заметим, что *негр* выступает в роли вспомогательного персонажа, которому не приписывается ни устойчивая речевая маска, ни постоянные черты характера.

090-Shmyrev_N:_Layout 1 18.05.2008 16:29 Page 585

СВОБОДНЫЕ РЕЧЕВЫЕ БАЗЫ ДАННЫХ VOXFORGE.ORG
VOXFORGE.ORG FREE SPEECH CORPUS

Шмырёв Н.В. *(nshmyrev@yandex.ru)*
НИИСИ РАН

В докладе обсуждаются проблемы, связанные с созданием свободных баз для систем синтеза и распознавания речи. Будут рассмотрены источники свободной речи, способы обработки информации и возникающие проблемы.

В связи с развитием устройств хранения и коммуникации современное оборудование позволяет накапливать и обрабатывать огромные массивы данных. Базы речи применяются при построении систем синтеза и распознавания речи, для оценки различных методик при тестировании приложений. Большинство современных баз собрано вручную, значительные ресурсы затрачены на их создание. Остро стоит вопрос автоматизации процесса сбора и обработки данных, вовлечения носителей языка в процесс записи. С одной стороны, значительные исследования посвящены попыткам организовать обработку с минимальным вмешательством исследователя, разработаны многие алгоритмы, позволяющие обойтись только незначительной предварительной обработкой. Например, анализ текстов [1] позволяет составить морфологическую модель языка только на основе словаря, без какой-либо дополнительной информации. С успехом подобные методы применяются и в области распознавания изображений. К сожалению, полностью автоматизировать процесс не всегда возможно, часто вмешательство человека все же требуется. Более того, на наш взгляд, невозможно собрать и обработать значительный объем речевой информации без привлечения большого числа пользователей. Наша задача – заинтересовать носителя языка, вовлечь его в процесс сбора данных, использовать его опыт и время при создании и проверке записей. Поэтому, актуальной является проблема организации процесса сбора данных, рассмотренная в данном докладе.

В этом докладе мы опишем методы, используемые для сбора речевых баз данных для систем распознавания в рамках проекта VoxForge.org [2], расскажем о текущих проблемах. Проект VoxForge.org посвящен сбору речевых данных для использования в системах распознавания речи. Мы собираем речевую базу на нескольких языках - английском, русском, немецком, итальянском, голландском. База распространяется свободно по лицензии GPL и содержит записи, разбитые на небольшие речевые отрезки в оригинальном формате записи и транскрипцию. Некоторые части содержат дополнительную разметку, например, разметку интонации, точную сегментацию и так далее, но это скорее исключение. Для каждого диктора в базе сохраняется возраст, пол. Диктор указывает свой диалект, которому, к сожалению, не стоит доверять. Объем собранной и обработанной речи для наиболее активных языков: английский – 40 часов речи более ста дикторов, русский – 10 часов речи более двухсот дикторов, немецкий и голландский языки – по 10 часов речи. В перспективе мы надеемся собрать значительно больший объем данных – до 140 речи часов для каждого языка. Скорость наполнения базы значительна – база английской речи пополняется примерно на 5 часов в месяц, более того, мы полагаем, что скорость пополнения будет расти. Растет и число поддерживаемых языков, в апреле 2007 года база содержала записи только на английском, в 2008 году – уже на 8 языках.

Проект распространяет фонетические словари, приложения для обработки данных, разметку речи на речевые сегменты, акустические модели для систем распознавания CMU Sphinx, HTK и Julius. Для Julius английская акустическая модель VoxForge.org является основной. Акустическая базы служат не только для создания систем распознавания, например, база русской речи используется в русском голосе msu_ru_nsh_clunits для синтезатора Festival [3]. Подобные проекты развиваются и в других областях, например, стоит отметить проект Freesound [4], посвященный сбору коллекции звуков.

Свободно распространяемая база имеет ряд преимуществ. Наиболее ценна она для свободных приложений. К сожалению, выбор свободных приложений в области речевых технологий невелик. Несмотря на наличие нескольких пакетов распознавания речи, в настоящий момент отсутствуют реализации некоторых необходимых компонентов речевых интерфейсов, таких как система управления диалогом. Мы надеемся, что наша работа стимулирует развитие свободных приложений. Свободная лицензия позволяет нам также снять и технические ограничения. Распределенное хранилище позволяет неограниченно расширять объем базы, не заботясь ни о сохранности данных, ни о производительности системы.

090-Shmyrev_N:_Layout 1 18.05.2008 16:29 Page 586

К счастью необходимо отметить, что для некоторых наиболее популярных языков источников речевых данных достаточно. Основной объем данных, собранных на ресурсе был прислан от обычных посетителей ресурса, но, в последнее время, появляются и другие источники данных. Посетитель ресурса может прислать свою запись следующими способами: по обычному или IP-телефону, записав данные на домашнем компьютере и прислав запись, посетив сайт и записав речь прямо с него. Последняя возможность наиболее важна для нас, как показывает практика запись с компьютера и телефона сложна для посетителей ресурса. Необходимость установки и настройки приложений отпугивает многих посетителей. Возможность записи речи прямо из web-браузера значительно увеличивает вклад посетителей в создание базы. Java-апплет не требует установки и настройки и позволяет прослушать записанную речь и загрузить речь в исходном формате на сервер.

При записи посетителю предлагается произнести текст, состоящий из случайно выбранных предложений из большого текстового корпуса, составленного из свободных текстов. Запись пользователя проверяется с помощью уже существующей речевой модели языка и, если проверка прошла успешно, сохраняется в базу.

По собранной базе периодически рассчитываются и обновляются акустические модели. Подготовка оптимизированных речевых моделей – сложный многоступенчатый процесс. На данный момент используется пакет HTK для расчета моделей для Julius и пакет SphinxTrain для расчета моделей для семейства приложений CMU Sphinx. В настоящий момент модели не оптимизируются, используются параметры расчета модели по умолчанию. Качество их, тем не менее удовлетворительное. Например, точность распознавания модели русского языка со словарем в 30 тысяч слов – порядка 70%. К сожалению, частота обновления ограничена скоростью расчета модели.

Нами используются и другие источники информации, которых последнее время становится все больше. Например, мы сотрудничаем с проектом по записи аудиокниг Librivox, поставляющем нам книги в оригинальном несжатом формате. Поступают предложения использовать фильмы с субтитрами, возможно даже использование записей официальных переговоров. Власти города Мизула в США в рамках программы OpenGoverment позволяют использовать запись обсуждений в городском совете. Эти записи уже содержат транскрипцию и могут быть использованы без значительных затруднений.

К сожалению, без дополнительного стимулирования не всегда возможно собрать требуемый объем речи. Мы используем и различные призы для участников, например, компания Voice2type предлагает приз наиболее активному пользователю нашего ресурса, приславшему запись речи с мобильного телефона. Тем не менее, проблема стимулирования остается открытой. Одной из самых интересных работ в направлении организации сбора данных и проведения вычислений с использованием человека является работа [5], рассматривающая использование человеческих ресурсов в сети. Остроумное использование web-технологий и соревнования между участниками процесса позволяет эффективно организовать распределенное вычисление, задействовать ресурсы пользователей и обработать огромный объем информации для дальнейших исследований в области распознавания изображений. Схема, позволяющая заинтересовать пользователя в записи, пока еще не разработана. В разработке находится модуль, позволяющий использовать собственную запись для адаптации акустической модели для пользователя, при этом будет необходимо записать некоторый небольшой текст. Такие записи мы будем использовать для дальнейшего улучшения общей модели. Нужно надеяться, таким образом удастся собрать действительно значительный объем речевых данных.

Другой, не менее важной проблемой является унификация и обработка собранных данных. Остановимся на проблеме первичной обработки данных. Тут возникают несколько проблем. Во-первых, формат данных не всегда соответствует текущим требованиям систем расчета моделей. Часто звук закодирован с потерей данных, например, в формат ogg или mp3. Возникает вопрос, возможно ли использовать такие данные для системы распознавания? К сожалению, внятных ответов на него пока не получено. Известно, что звук, декодированный из mp3 отрицательно сказывается на качестве акустической модели, в тоже время некоторые исследователи в устной беседе утверждают, что с помощью моделей на основе речи, сжатой в формате mp3 можно отлично декодировать mp3 речь. Возможно ли смешивать данные с различной степенью сжатия, и как это сказывается на качестве распознавания, пока неизвестно. Нужно надеяться, современные методы извлечения параметров речи при расчете моделей позволят решить эту проблему.

Во-вторых, интересным направлением исследования является проблема выделения речевых отрезков и проверки большой записи, например, разбиения аудиокниги на небольшие куски, пригодные для расчета параметров акустической модели. Задача разбиения большого звукового файла имеет и самостоятельное значение. Она находит применение в системах обучения языку, где временные метки используются приложением для отображения синхронного перевода. Необходимо заметить, что в последнее время значительное количество работ позволило добиться успеха в этой области. Например, система [6] позволяет обработать аудиокнигу и получить базу речевых отрезков и транскрипцию. В настоящее время мы применяем разбиение с помощью выравнивания по существующий модели и тексту. Наиболее сложной проблемой является наполнение словаря

090-Shmyrev_N:_Layout 1 18.05.2008 16:29 Page 587

неизвестными словами из текста. Для наполнения словаря используется система автоматической транскрипции, обученная по существующим словарям. Наиболее точные системы, основанные на мультиграммных поточных моделях [7] обеспечивают точность транскрипции порядка 70%. Остальные 30% приходится корректировать вручную.

К сожалению, присутствуют и более серьезные проблемы, порожденные самим способом сбора данных. К сожалению, несмотря на значительные объемы хранимой информации, тяжело получить хорошо сбалансированную базу данных. Объем данных из некоторых источников доходит до 10 часов, часто это речь всего одного диктора. Конечно, таким базам тоже можно найти применение. Например, их можно использовать для создания систем высококачественного синтезатора речи. Не ясно, насколько такая несбалансированная база будет полезна в исследованиях по распознаванию речи. Мы надеемся, что возможно будет выделить некоторую сбалансированную часть базы.

Стилевое наполнение базы тоже не оптимально. Для некоторых задач, например, задачи поиска в звуковых файлах или в задаче транскрибирования видео более полезны записи повествовательной речи. Важная задача создания речевого интерфейса требует базы данных совсем другого стиля, базы, составленной в основном из диалогов и спонтанной речи. Увеличить наполнение диалоговой составляющей базы – задача на ближайшее будущее.

Мы описали работу, которая ведется в направлении создания системы распознавания речи на основе свободных речевых данных. Нужно признать, значительные аспекты еще не проработаны. Не исследованы проблемы построения акустической модели нескольких языков, создания оптимального текста и условий для записи. К счастью, круг проблем еще очень широк.

Список литературы

1. Mathias Creutz, Krista Lagus. Unsupervised Models for Morpheme Segmentation and Morphology Learning // ACM Transactions on Speech and Language Processing, Volume 4, Issue 1, Article 3, January 2007.

2. VoxForge project // http://voxforge.org

3. Русский голос для Festival // http://festlang.berlios.de/russian.html

4. FreeSound project // http://freesound.iua.upf.edu

5. Luis von Ahn, Laura Dabbish, Labeling images with a computer game. // Proceedings of the SIGCHI conference on Human factors in computing systems, 2004.

6. Prahallad K., Toth A., Black A. Automatic Building of Synthetic Voices from Large Multi-Paragraph Speech Databases // Interspeech 2007, Antwerp, Belgium.

7. Bisany M., Ney H. Joint-sequence models for grapheme-to-phoneme conversion. // Speech Communication,Volume 50, Issue 5, May 2008.

091-Yagunova_E:_Layout 1 18.05.2008 16:30 Page 588

НАБОР ОПОРНЫХ СЛОВ КАК ВИД СВЕРТКИ ТЕКСТА (В СОПОСТАВЛЕНИИ С НАБОРОМ КЛЮЧЕВЫХ СЛОВ)*

SET OF RECOGNIZABLE WORDS AS COMPRESSION TEXTS (WITH COMPARISON OF KEY-WORD SET)

Ягунова Е.В. *(iagounova_elena@mail.ru)*
Санкт-Петербургский государственный университет

В докладе рассматриваются основные характеристики набора опорных слов как свертки текста. Под опорными понимаются наиболее распознаваемые слова при восприятии текста в шуме. Наши данные подтверждают гипотезу о том, что существенное значение для смыслового структурирования текста имеет функциональный стиль и степень динамичности текста.

1. Вводные замечания

Свертка текста – это упорядоченный набор слов, сопоставленный исходному тексту и в определенных условиях выступающий в качестве его «представителя». Порядок слов в свертке обычно воспроизводит порядок их появления в исходном тексте (хотя возможны и отклонения), сами слова отбираются по принципам, определяемым целью свертывания текста (см. об этом ниже). В настоящей работе анализируются два типа сверток. Первый – это наиболее традиционная разновидность – набор ключевых слов (НКС), в котором задан порядок – порядок введения ключевых слов в текст. Ключевые слова отражают тему текста, а упорядоченность слов в НКС активирует ассоциативные связи, необходимые для существования любого текста (что проявляется в возможности развертывания НКС в цельный и связный текст) (Мурзин, Штерн 1991; Сахарный и др. 1984).

Наибольший интерес для нас в настоящей работе представляет другой вид сверток – наборы опорных слов (НОС). Под опорными в настоящем докладе понимаются слова, распознающиеся не менее чем 30% испытуемых при восприятии текста в шуме при соотношении сигнал/шум 0 дБ. Распознаваемость этих слов определяется влиянием как фонетических, так и внефонетических признаков. Наиболее разборчивыми могли стать те слова, которые идентифицировались за счет фонетической информации и/или за счет высокой контекстной предсказуемости. Можно предположить, что эти слова несут особую смысловую нагрузку, которой и обусловлена высокая распознаваемость именно этих слов (в частности, соответствующая фонетическая оформленность и/или контекстная предсказуемость опорных слов)[1].

В качестве исследуемого материала было выбрано два переводных текста на русском языке:

1. Отрывок из официальной публикации «Закон об иностранных инвестициях во Вьетнаме и нормативные акты, изданные на его основе» (в дальнейшем «деловой текст»);

2. Отрывок сюжетной художественной прозы Нам Као «Ти Фео» с элементами диалога (в дальнейшем «художественный текст»).

Художественный текст является динамическим, т.е. характеризуется сменой ситуаций. Композиционная структура художественного текста соответствует традиционной схеме нарратива: преамбула (ориентация), завязка, развитие сюжета, развязка (по Чейфу). Деловой текст (текст-предписание) является статическим.

В эксперименте по восприятию текста в шуме приняло участие 40 испытуемых. В инструкции испытуемым предлагалось прослушать текст удобными порциями, останавливаясь с помощью клавиши «пауза» и записывая каждый услышанный фрагмент. Текст можно было слушать один раз, не возвращаясь назад.

В настоящем исследовании восприятие текста носителями языка осуществляется – в любых экспериментах – в условиях ограничения на «базу знаний» слушающего, т.к. испытуемыми выступают люди, далекие от предметной области экономики и делопроизводства (деловой текст) и реалий вьетнамской жизни (художественный текст).

При выборе стратегий восприятия особую роль должно играть начало текста как «площадка», в рамках которой происходит подстройка под особенности воспринимаемого текста. Соответственно рассматриваются противопоставления:

* Работа выполнена при частичном финансировании РГНФ (проект №07-04-00161а)

[1] При восприятии осмысленного текста разделить вклад фонетической и внефонетической информации невозможно. В формат доклад не вошел дополнительный эксперимент по восприятию аналогичного асемантического текста (см., напр., Ягунова 2006), в котором опорными являлись слова, хорошо распознаваемые в силу своих собственно фонетических свойств.

091-Yagunova_E:_Layout 1 18.05.2008 16:30 Page 589

- начальный фрагмент делового текста vs. весь деловой текст
- начальный фрагмент художественного текста[2] vs. весь художественный текст

КС определялись – по традиционно используемой методике – в ходе вспомогательного эксперимента, в котором испытуемым предлагалось следовать следующей инструкции: «Прослушайте текст. Подумайте над его содержанием. Выпишите[3] из текста 10-15 слов, наиболее важных с точки зрения его содержания». Испытуемые заполняли анкеты ***после*** окончания прослушивания текста. Следовательно, НКС, по-видимому, отражает смысловую структуру текста как целостного объекта.

Распределение в тексте наиболее разборчивых слов – НОС – соответствует, вероятно, «следам» восприятия и понимания текста ***в текущем режиме времени***[4], когда смысловая структура извлекается не только из текста целиком, но и из его структурных составляющих.

Каково соотношение двух наборов – НОС и НКС?

Оба вида сверток отражают, тем или иным образом, смысловую структуру текста. Казалось бы, КС как смысловые вехи текста должны формироваться из наиболее частотных в тексте слов и характеризоваться (1) наибольшей разборчивостью (т.е. быть одновременно и опорными словами); (2) наилучшей предсказуемостью на конечном фрагменте текста (в результате подстройки слушающего под особенности текста). Однако реальная ситуация оказывается сложнее. Как показали предыдущие исследования автора (Ягунова 2007), эти характеристики КС (по сравнению с прочими, неключевыми словами (неКС)) обнаруживают значимую зависимость от **коммуникативной ситуации** – отфункционального стиля текста и экспериментального режима прослушивания.

В деловом тексте частотность словоупотреблений (и конструкций) по тексту высокая, порядок следования частотных словоупотреблений соответствует потенциальным путям формирования КС и важен для обеспечения связности текста. В художественном тексте частотность знаменательных словоупотреблений существенно ниже, чем в деловом тексте, т.е. частотных по тексту слов не может быть достаточно для формирования НКС.

В режиме восприятия текста в шуме КС делового текста распознавались существенно лучше, чем неКС, для художественного текста, напротив, неКС распознавались значимо лучше, чем КС. По мере продвижения слушающего по тексту (от начального к конечному фрагменту текста) происходит улучшение распознаваемости КС для обоих текстов.

В парадигме гештальтной психологии КС могут рассматриваться как фигура, а неКС – как фон. В случае рассматриваемого художественного текста, характеризующегося сменой ситуаций, соотношение «фигура – фон» может меняться от ситуации к ситуации (т.е. для одних ситуаций фигура плохо выделяется на фоне, для других – хорошо). Можно предположить для этого художественного текста наличие несоответствий между КС, выделяемыми после прослушивания текста и опорными словами, выделяемыми во время прослушивания текста в текущем времени. В случае статического – делового – текста соотношение «фигура – фон» сравнительно стабильно. Соответствие между КС и опорными словами, область их пересечения, таким образом, должны быть существенно выше для делового текста.

Далее будет рассмотрена ***гипотеза*** о том, что существенное значение для смыслового структурирования и формирования сверток текста имеют признаки «функциональный стиль текста» и «степень динамичности текста».

Функциональный стиль и степень динамичности текста могут рассматриваться как взаимодополняющие друг друга признаки. Проверка гипотезы осуществляется на основании сопоставления двух типов сверток (для КС и опорных слов), показывающего (1) как часто опорные слова являются одновременно и КС; (2) представленность основных частеречных классов в рассматриваемых наборах.

В соответствии со сформулированной гипотезой для статичного делового текста соответствие между НКС и НОС должно быть значительно выше, чем для динамического художественного текста, содержащего несколько последовательных ситуаций. Начальный фрагмент художественного текста, включающий лишь преамбулу и завязку, описывает меньшее число ситуаций и характеризуется меньшей динамичностью, чем весь художественный текст. Следовательно, согласно выдвинутой гипотезе, соответствие между двумя типами сверток для начального фрагмента художественного текста должно быть выше, чем для всего художественного текста.

---

[2] Начальный фрагмент соответствует примерно трети от каждого текста. Для художественного текста он соотносится с двумя композиционными компонентами: преамбула и завязка.

[3] Использование глагола «выписать» в инструкции продиктовано желанием (1) акцентировать испытуемых на задачу представления в своих списках слов, действительно присутствующих в тексте и (2) следовать традиции для возможности сопоставления, напр., НКС, полученных на материале устных и письменных текстов. Ключевые слова записывались испытуемыми ***после*** прослушивания. Эксперимент был проведен в двух вариантах: при предъявлении всего текста (КС, НКС) и его начального фрагмента (КСнач, НКСнач). В разных экспериментах (или сериях) участвовали разные испытуемые.

[4] Текст испытуемые могли слушать только один раз, удобными порциями, не возвращаясь назад.

091-Yagunova_E:_Layout 1 18.05.2008 16:30 Page 590

2. Сопоставление наборов ключевых слов и наборов опорных слов

В настоящей работе в качестве фрагмента, на котором предположительно происходит подстройка слушающего под смысловые особенности текста, был выбран начальный фрагмент. Этот фрагмент соответствует трети от всего текста. Следовательно, при прослушивании начального фрагмента в наименьшей степени задействовано возможное противоречие между собственно смысловой структурой всего текста и его структурных компонентов.

В табл. 1а и 1б приведены свертки трех видов: НОС, НКС и НКСнач. Слова в этих наборах соответствуют словоупотреблениям из текста, а НКСнач обозначает набор ключевых слов, выделенных по традиционной методике на основании эксперимента, в котором предъявлялся только начальный фрагмент делового и художественного текстов.

НОС	НКСнач	НКС
стороны	**стороны**	**стороны**
участвующие		
	договора	**договора**
	о деловом	о деловом
	сотрудничестве	**сотрудничестве**
	и предприятия	и предприятия
с иностранным	**с иностранным**	**с иностранным**
капиталом	**капиталом**	**капиталом**
имеют		
право		
самостоятельно		
устанавливать		
программы	**программы**	
	хозяйственной	
	показатели	
		вьетнамскими
	государственными	государственными
	органами	
	экономическом	
		разрешения
	на вложение	
	инвестиций	
	на деловое	на деловое
	сотрудничество	**сотрудничество**
стороны	**стороны**	**стороны**
договора	**договора**	**договора**
	о деловом	о деловом
сотрудничестве	**сотрудничестве**	**сотрудничестве**
	предприятий	предприятий
с иностранным	**с иностранным**	**с иностранным**
капиталом	**капиталом**	**капиталом**
	показатели	

Таблица 1а. Три вида наборов для начального фрагмента делового текста

НОС	НКСнач	НКС
когда	стороны	стороны
подобные		
мысли	**мысли**	
трудно		

091-Yagunova_E:_Layout 1 18.05.2008 16:30 Page 591

настроение	**настроение**	
	и выдержку	
		Ти Фео
		клянчить
	на выпивку	
	советник	советник
	бросить	
	хао	хао
пора бы		
знать		
свое		
место	**место**	
	не банк	
	швырнул	
	монету	
		и крикнул
		притащился
	забирай	
	и катись	
	духу	
	научись	
	жить	
	честно	

Таблица 1б. Три вида наборов для начального фрагмента художественного текста

Примечания к табл. 1а и 1б: словоупотребления упорядочены в соответствием с порядком введения в исходный текст[5]

Деловой текст

Слова из НОС в 86% случаев принадлежат НКСнач и в 79% случаев НКС[6]. Таким образом, опорные слова в подавляющем числе случаев являются и ключевыми словами.

Ключевые словоупотребления – НКС – в 63% случаев являются опорными словами, НКСнач лишь в 48% случаев являются опорными.

Художественный текст

Для НОС художественного текста лишь три слова (*мысли*, *настроение, место*) находятся на пересечении с НКСнач (33% опорных слов являются ключевыми и 18% ключевых являются опорными); пересечения НОС и НКС отсутствуют. Как и следовало ожидать, наибольший объем имеет НКСнач (см. табл. 1б).

Данных о частеречных характеристиках словоупотреблений из рассматриваемых наборов приведены в таблицах 1в и 1г).

текст	часть речи	НОС	НКСнач	НКС
деловой	существительные	57	68	63
	глаголы	21	0	0
	прилагательные	14	32	37
художественный	существительные	33	59	50
	глаголы	11	35	50
	прилагательные[7]	56	0	0

Таблица 1в. Представленность основных частеречных классов в рассматриваемых свертках (начальный фрагмент) (%)

[5] Пересечения между НОС и НКС (или НКСнач) в рамках таблицы выделены полужирным шрифтом; наличие такого пересечения фиксировалось, если одна и та же лексема присутствовала в этих наборах вне зависимости от порядка следования (ср. лексему «договор»)

[6] О функциональном подобии структур «КС vs. неКС» и «КС vs. неКС» для начального фрагмента делового текста см. (Ягунова 2007).

[7] Для НОС художественного текста этот частеречный класс включает не только прилагательные, но и наречия, и предикативы

091-Yagunova_E:_Layout 1 18.05.2008 16:30 Page 592

текст	часть речи	НОС	НКС
деловой	существительные	43	65
	глаголы	13	4
	прилагательные	14	37
художественный	существительные	19	70
	глаголы	19	25
	прилагательные	62	5

Таблица 1г. Представленность основных частеречных классов в рассматриваемых наборах (весь текст) (%)

Результаты анализа НОС (в сопоставлении с НКС) для делового и художественного текстов подтверждают выдвинутую гипотезу.

✓ Наборы ключевых слов (НКС и НКСнач) существенно различны для делового и художественного текстов: они содержат больше глаголов для художественного текста, что отражает бóльшую степень его динамичности по сравнению с деловым текстом.

✓ Пересечение НОС и НКС существенно выше для делового текста по сравнению с художественным. Для делового текста подавляющее число опорных слов являются ключевыми (86% vs. 33%) для начального фрагмента, более половины опорных слов являются ключевыми при рассмотрении целого текста (52% vs. 13%).

✓ Предположение о том, что смысловая структура, заложенная в НОС, является более динамичной, чем смысловая структура, представленная в виде НКС, подтверждается только для целых текстов. Для них это проявляется в увеличении доли глаголов для делового текста и глаголов, «дополненных» предикатной лексикой для художественного текста.

✓ Для начального фрагмента текста на основании анализа НОС сделать заключение о степени динамичности исходного текста, по-видимому, невозможно. Для начального фрагмента в случае НОС – в отличие от наборов ключевых слов (НКС и НКСнач) – нивелируются различия между степенью динамичности исходного текста.

3. Восстановление текста на основе наборов опорных слов

Признаки «функциональный стиль» и «степень динамичности», заложенные в НОС, необходимо проследить также через восстановление внутритекстовых ассоциативных связей между словами (в том числе связей в разных составляющих: синтагмах, фразах, ситуациях и т.д.), т.е. через эксперимент по восстановлению текстов на основе НОС.

В основу описываемого далее эксперимента легли следующие ***гипотезы***:

• опора на НОС позволяет осуществить построение целостного связного текста;

• НОС задают функциональный стиль развертываемого текста;

• НОС отражают предметную (тематическую) область развертываемого текста;

• развертывание НОС позволяет определить позиции (контексты) слов и/или конструкций, обладающих максимальной предсказуемостью.

Проверка трех последних гипотез возможна лишь в том случае, если первая гипотеза верна, т.к. исследованию подлежат именно восстановленные испытуемыми тексты.

Эксперимент проводился в письменно-письменной форме. Опорные слова представляли собой фонетические слова, т.е. знаменательные словоформы с клитиками (предлогами, союзами, частицами). Анкета для испытуемых включала два НОС, записанных в столбцы под названием «текст 1» и «текст 2», и письменной инструкции. Столбцы соответствовали наборам для делового («текст 1») и художественного («текст 2») текстов одного экспериментального режима. В докладе рассматриваются некоторые результаты двух серий эксперимента: для начального фрагмента текста и всего текста. В письменной инструкции было указано: «*Перед Вами* ***последовательность*** *слов, извлеченных из текста. Попробуйте на их основе восстановить текст*», письменная инструкция дополнялась устным требованием сохранять грамматическую форму слов, вводить слова в текст в указанной последовательности и пожеланием минимизирования литературных и «философских» фантазий. Время на выполнение задания не ограничивалось.

По условиям эксперимента испытуемые не владели предметной областью экономики, делопроизводства (деловой текст), были незнакомы с реалиями жизни во вьетнамской деревне (художественный текст). Это требование позволило рассматривать процедуры понимания текста в условиях ограничений на «базу знаний» адресата.

091-Yagunova_E:_Layout 1 18.05.2008 16:30 Page 593

Анкеты, не удовлетворяющие требованиям инструкции, далее не рассматривались. В результате для серии «начальный фрагмент текста» было получено более 30 анкет, для серии «весь текст» – 20 анкет. Каждый испытуемый участвовал только в одной из серий эксперимента.

Определение того, принадлежат ли восстановленные тексты тому же функциональному стилю, производилось на основании двух критериев:

- экспертная оценка:
  - заключение эксперта о принадлежности восстановленного текста к данному функциональному стилю (деловому или художественному),
  - степень статичности vs. динамичности смены описываемых ситуаций (как дополнительный признак);
- количественные (формальные) критерии:
  - коэффициент лексического разнообразия текстов (КЛР)[8], что отражает степень разнообразия лексических средств при построении текста и соотносится с функциональным стилем текста,
  - длина текстов в словах (как дополнительный признак).

Предполагается, что разнообразие лексики (высокий КЛР) характеризует динамический (художественный) текст, а клишированность (низкий КЛР) – статический деловой текст.

4. Результаты эксперимента. Выводы

На основании данных, полученных в результате эксперимента по восстановлению текста, можно сформулировать следующие выводы.

	деловой		художественный	
	НОС для всего текста	НОС для начального фрагмента	НОС для всего текста	НОС для начального фрагмента
длина текстов (в с/у)	124	51	118	64
КЛР для словоформ	0,62	0,81	0,80	0,86
КЛР для лексем	0,51	0,74	0,72	0,79

Таблица 2. Средние значения КЛР и длины восстановленных текстов

Деловой текст

• НОС для всего текста

1. Все восстановленные на основе НОС тексты – как и исходный текст – относятся к деловому функциональному стилю. Большинство развернутых текстов можно отнести к жанру нормативных актов. Все развернутые тексты – как и исходный текст – относятся к статическому варианту: как правило, предписание, регулирующее положение дел, или – в ряде случаев – описание некоторого положения дел.

2. Высокая степень повторяемости словоформ (коэффициент лексического разнообразия КЛР 0,46) в предъявляемом НОС провоцирует высокую повторяемость слов в восстанавливаемых испытуемыми текстах (см. табл. 2).

3. При развертывании текста испытуемыми восстанавливались КС (как присутствующие, так и отсутствующие в НОС), т.к. они соответствовали предметной области данного текста. Подавляющее число используемых слов относятся к деловому функциональному стилю[9].

• НОС для начального фрагмента

1. Все развернутые тексты – как и исходный текст – относятся к деловому функциональному стилю. Большинство развернутых текстов можно отнести к жанру нормативного акта.

2. Значения КЛР для начального фрагмента исходного делового текста показывает сравнительно высокую степень повторяемости лексических единиц, однако степень клишированности исходного делового текста на всем тексте существенно выше. Клишированность текста лишь отчасти задается через НОС (КЛР 0,82). КЛР для лексем восстановленных текстов в среднем близок к соответствующему показателю исходного текста. КЛР для словоформ восстановленных текстов в среднем ниже, чем для исходного текста (для этого фрагмента исходного тек-

[8] КЛР подсчитывается следующим образом: в числителе – число разных единиц, в знаменателе – общее число рассматриваемых единиц в тексте.

При отнесении лексики к деловому функциональному стилю использовался сделанный С.А. Шаровым на основе НКРЯ прототип частотных списков по жанрам (http://corpus.leeds.ac.uk/serge/frqlist/), в котором представлен «Частотный словарь административных текстов, значимая лексика»

091-Yagunova_E:_Layout 1 18.05.2008 16:30 Page 594

ста КЛР 0,40 и 0,61, соответственно).

3. Различия между КЛР между восстановленными деловыми текстами на основании «НОС для всего текста vs. НОС для начального фрагмента» существенны как для словоформ, так и для лексем (см. табл. 2). По-видимому, это различие соотносится с важностью текстового «окна сверки» для восприятия делового текста (Ягунова 2007).

Художественный текст

• НОС для всего текста

1. Все развернутые тексты – как и исходный текст – относятся к художественному функциональному стилю. Они характеризуются динамичностью. Обычно восстанавливается сюжет с двумя действующими лицами, в тексте присутствуют диалоговые фрагменты; характеристики действующих лиц (как и место действия) могут быть самыми разными. Общее содержание развернутых текстов может быть охарактеризовано как реализации сценария «конфликт» (сопровождающийся каким-либо требованием).

2. Художественный текст с динамично развивающимся сюжетом характеризуется разнообразием знаменательной лексики. Повторяемыми являются, главным образом, местоименная лексика, в какой-то степени – наименования действующих лиц.

3. КЛР для словоформ и лексем восстановленных текстов несколько выше, чем для исходного текста. Это незначительное расхождение, вероятно, связано с индивидуальными стратегиями порождения текста (различаются богатством сюжетной линии, использованием синонимических ресурсов и местоименной лексики).

• НОС для начального фрагмента

1. Все развернутые тексты – как и исходный текст – относятся к художественному функциональному стилю. Большинство развернутых художественных текстов характеризуется статичностью (отсутствием смены ситуаций). Как правило, при разворачивании НОС для начального фрагмента текста не восстанавливается сюжет с двумя действующими лицами, в тексте отсутствуют диалоговые фрагменты.

2. Степень разнообразия лексем может быть связана со степенью статичности текста: восстановленные тексты являются более статичными, чем исходные, и КЛР для лексем восстановленных текстов ниже, чем для исходного текста. Возможным подтверждением статичности восстанавливаемого фрагмента художественного текста служит то, что КЛР для лексем восстановленных художественных и деловых текстов (начальные фрагменты) различаются лишь на уровне тенденции (см. табл. 2).

Главным результатом является подтверждение гипотезы о том, что НОС задают функциональный стиль текстов, восстанавливающихся испытуемыми в эксперименте. Значимость различий между КЛР (для словоформ и лексем) восстановленных (по НОС для всего текста) деловых и художественных текстов является количественным показателем принадлежности текстов к разным функциональным стилям.

Полученные данные не противоречат гипотезе о том, что НОС (для начального фрагмента текст) задают функциональный стиль тех текстов, что восстанавливаются испытуемыми в эксперименте. То, что восстановленные тексты принадлежат тому же функциональному стилю, что и исходные, подтверждается качественными и – отчасти – количественными критериями.

Результаты эксперимента по развертыванию НОС подтверждают выдвинутую гипотезу о том, что существенное значение для смыслового структурирования текста имеет функциональный стиль и степень динамичности текста. Наиболее ярко взаимодействие факторов «***функциональный стиль***» и «***степень динамичности текста***» проявляется именно в случае наличия противоречия этих факторов.

Список литературы

1. Мурзин Л.Н., Штерн А.С. Текст и его восприятие. Свердловск, 1991

2. Сахарный Л.В., Сибирский С.А., Штерн А.С. Набор ключевых слов как текст // Психолого-педагогические и лингвистические проблемы исследования текста – Пермь, 1984

3. Ягунова Е.В. Мелодические признаки и опорные элементы при восприятии текста // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2006» (Бекасово, 31 мая – 4 июня 2006 г.) / Под ред. И.М. Кобозевой, А.С. Нариньяни, В.П. Селегея.– М.: Наука, 2005

4. Ягунова Е.В. Коммуникативная и смысловая структуры текста и его восприятие // Вопросы языкознания. 2007, №6

092-Yanko_T:_Layout 1 18.05.2008 16:57 Page 595

ПРОСОДИЯ В ТОЛКОВОМ СЛОВАРЕ И СЛОВАРЬ УНИКАЛЬНЫХ ПРОСОДИЙ[1]

PROSODY IN A DICTIONARY, AND A DICTIONARY OF PROSODIC IDIOMS

Янко Т.Е. *(tanya_yanko@list.ru)*
Институт языкознания РАН

Проблема представления просодической информации в словаре языка разбивается на две подзадачи: объяснение ограниченных коммуникативных и просодических возможностей лексем с помощью толкования их значений и анализа функций; инвентаризация идиоматических просодий и соответствующих иллокутивных сил.

В современных словарях лексикографический «портрет» слова включает сведения о его индивидуальных коммуникативных и просодических свойствах (Апресян 1988). В работах Т.В Булыгиной анализируются слова *мало*, *несколько*, *редко*, *вечно*, в работах Е.В Падучевой (1997) – *редкий*, *давно*, *долго*, И.М. Богуславского (1996: 63) – *задолго* и *незадолго*, Ф.И. Панкова (1996, 2005, 2006) – *скоро*, *близко* и др. Так, Т.В. Булыгина обратила внимание на то, что в глагольной группе *принес много книг* ремой может быть и словоформа *книг* с рематическим акцентом на *книг* (*принес много* ***книг***) и словоформа *много* (*принес* ***много*** *книг*), а в *принес мало книг* – только словоформа *мало*: *принес* ***мало*** *книг* (ср. **принес мало* ***книг***). В работе (Булыгина, Шмелев 1997: 200-207) рематичность *редко* и *мало* объясняется семантикой несоответствия норме: *редко* и *мало* – меньше, чем предполагается нормой, а ее несоблюдение не согласуется с коммуникативной ролью темы, которой свойственно отображать известное, освоенное, привычное и нормальное. Толкования слов могут иметь пересечения с толкованиями коммуникативных значений, что приводит к ограниченной возможности соответствующих лексем выступать в тех коммуникативных ролях, которые противоречат этим значениям. Кроме того, толкованиям коммуникативных значений, таких, как темы и ремы, могут противоречить не только толкования слов, но и значения, полученные из толкований путем логического вывода. В результате выделяются слова с т.н. «рематической полярностью» (*редко*, *мало*), «тематической полярностью» (*теперь*, *наконец*), слова контраста (*только*, *именно*), слова верификации (*правда*, *вправду*), слова эмфазы (*сам*, *даже*) и т.н. неконтрастоспособные конструкции, чье толкование мешает образованию релевантного множества, из которого при контрасте производится выбор. В (Янко 2001: 306-325) показано, что пропозициональное дополнение глагола *знать* в силу онтологических свойств знания не может вступать в отношение контраста, потому что один объект знания, как правило, не противопоставлен другому. Если мы едим грушу, то весьма вероятно, что одновременно мы не едим также и яблоко, поэтому разумно говорить о том, что мы едим грушу, а не яблоко. Между тем, если нам известно о факте P, это никак не мешает нам знать и о факте Q: одно знание не исключает другого. Ср. *ем не P*, *а Q*, *люблю не P*, *а Q*, но не **знаю не что P*, *а что Q*. Результаты, которые касаются коммуникативных ограничений, вытекающих из толкований, были получены нами ранее и опубликованы (Янко 2001), поэтому здесь в разделе 1 приводится только один пример, иллюстрирующий соответствующую проблематику.

Свойства слов – это только одна сторона проблемы представления просодической информации в лексиконе языка, потому что просодические структуры тоже могут быть объектами словарного описания, если соответствующие означаемые не сводятся к композициям других коммуникативных значений и сами не вступают в композиции. Ср., с одной стороны, значения темы, ремы и контраста, которые образуют стандартные композиции – контрастные темы и ремы – и, с другой стороны, рассматриваемую ниже иллокуцию зова адресата в условиях поиска, которая не вступает в композиции с другими значениями и, соответственно, требует словарной фиксации. Более того, план выражения коммуникативных значений тоже может обнаруживать идиоматичную просодию, которая характеризуется такими признаками, как: 1) нечленимость цепочки изменений частоты основного тона на известные элементарные единицы типа интонационных конструкций по Е.А. Брызгуновой[2] или на еще более мелкие единицы; 2) нестандартный тип связи частотных изменений с сегментной структурой, например, независимость частотных изменений от распределения ударных и безударных слогов, при том что неидиоматич-

[1] Работа над темой финансируется Российским Гсударственным гуманитарным фондом, проект N08-04-00165а.
[2] Об интонационных конструкциях по Е.А. Брызгуновой см. Русская грамматика 1982: 103-118.

092-Yanko_T:_Layout 1 18.05.2008 16:57 Page 596

ные структуры реализуют релевантные движения тона с жесткой привязкой к ударным слогам слова-акцентоносителя, 3) нестандартная связь просодических структур со словами-акцентоносителями, которые выбирались бы по определенным синтаксически обусловленным правилам, как это происходит при формировании стандартных структур[3]. Имеются и другие свидетельства идиоматического оформления речевых актов, когда принцип композициональности не действует и специфическая мелодика ложится на сегментный материал, как музыка на слова песни, as when chanted по выражению знаменитой исследовательницы английской интонации Дж. Пьерхамберт (Pierrehumbert 1980: 87-88).

В докладе к традиционной задаче анализа уникальных коммуникативных и просодических свойств слов и конструкций добавляются следующие задачи: 1) установление связи идиоматичного просодического оформления лексем не только с толкованием, но и с функцией в тексте (в связи с этим ниже будет рассмотрено нестандартное поведение притяжательных местоимений *мой*, *твой*, *наш* и *ваш* в структуре обращений *Ваша честь*! *Сын мой*! *Рыбка моя*! *Птичка моя*! *Неподкупный вы наш*!), и 2) разработка словаря просодических идиом.

В разделе 1 говорится о словарных коммуникативных и просодических ограничениях в употреблении слов и конструкций. В разделе 2 на примере речевого акта обращения рассматриваются типы речевых актов, имеющих идиоматичное выражение. Раздел 3 посвящен уникальным просодическим свойствам местоимений *мой*, *твой*, *ваш* и *наш* в составе обращений.

1. Просодическая информация о слове

В данном подразделе мы приведем пример обусловленности коммуникативной роли слова и соответствующей просодии толкованием. Обратимся к слову *давно*. Е.В. Падучева показала, что это слово не бывает в предложении нормальной темой, а в контексте общефактического значения несовершенного вида *давно* не только «антитематично», но еще и «рематично». Оно служит акцентоносителем ремы: *Папа покупал эти часы* ***давно*** (но не **Папа давно покупал эти* ***часы***) (Падучева 1997). Впоследствии в наших работах (Янко 2001: 255-269) и (Yanko 2003) было показано, что рематическая полярность у слова *давно* проявляется не только в контексте общефактического, но и в контексте всех других видо-временных глагольных форм, которые обозначают событие, ушедшее в прошлое, т.е. таких, которые поддерживают семантику разобщенности с моментом речи, заключенную в *давно*: *Это произошло/происходило в нашем городе* ***давно*** (но не **Это давно произошло в нашем* ***городе***). Между тем в контексте глагольных времен, которые обозначают время, включающее момент речи, *давно* – не обязательно рема: *Вася давно* ***спит***; *Вася давно* ***пришел***. Рематическая полярность *давно*, которая выражена соответствующей реме просодией, объясняется семантикой временно́й удаленности события от момента речи: то, что происходило давно, далеко от момента речи, чуждо ему, поэтому соответствующие лексемы не могут служить темой – исходной точкой для совершения речевого акта. Это подтверждается тем, что наречие *недавно* в контекстах, которые для *давно* оказываются контекстами исключительной рематичности, не обязательно служит ремой: *Это недавно произошло в нашем* ***городе.***

Возникает вопрос, почему выражение *давным давно* практически с той же семантикой, что и у *давно*, – это классическая тема предложения (*Давным давно жил-был царь*)? Мы предполагаем, что у *давным давно* есть дополнительный компонент в значении, который переориентирует полярность *давным давно* по сравнению с *давно*. Таким значением оказывается экзистенциальная квантификация, показатели которой имеют склонность входить в тему предложения (Падучева 1974:87-88). В качестве предварительного толкования *давным давно* мы предлагаем следующую формулировку:

Давным давно P ≈ ‘Существует событие P такое, что время, когда произошло P, имело место до момента речи (или другой точки отсчета времени)’.

Между тем для предложений типа *P давно* предлагается следующее толкование:

P давно ≈ ‘Время, когда произошло событие P, имело место до момента речи (или другой точки отсчета времени)’.

Толкования предложений с *давно* и *давным давно* различаются – в частности – компонентом ‘существует’. Аналогичные толкования можно предложить для пар слов с близкими значениями, но различной коммуникативной ролью и, соответственно, с различной просодией: рематичного *редко* и универсального *изредка*, рематичного *мало* и универсальных *немного* и *несколько*. Анализ других слов и конструкций темы, ремы, конраста, эмфазы, верификации и неконстрастоспособных слов см. в Янко 2001: 231-337 и в цитированной там литературе.

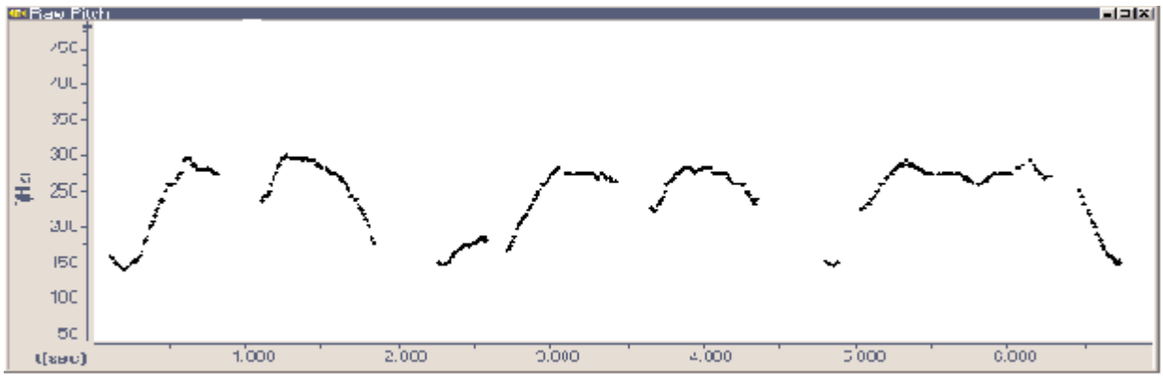
[3] Примеры идиоматических просодий читатель найдет ниже в разделе 2.

092-Yanko_T:_Layout 1 18.05.2008 16:57 Page 597

2. Речевой акт как единица просодического словаря. Речевой акт обращения

Уникальные иллокуции представляют собой открытое множество, и для многих из них интонация служит основным средством выражения. Для представления здесь уникальных иллокуций в качестве примера рассмотрим два типа обращений: зов в условиях существенной удаленности говорящего от слушающего и зов в условиях поиска.

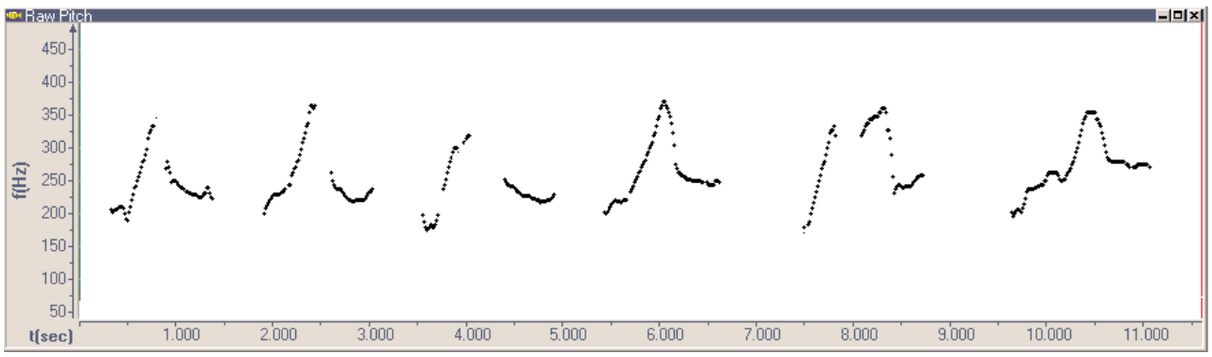
Обратимся к ситуации, в которой слушающий существенно удален от говорящего, но говорящий, тем не менее, видит слушающего и считает, что может «докричаться» до него. Тонограмма ниже показывает, что в этом случае, независимо от сегментного материала, на имени адресата фиксируется одна и та же акцентная кривая, которая занимает примерно одинаковое время звучания независимо от длины слова.



Ва-ася-я *Ната-аша-а* *Джо-он*

Сравнительный анализ данного акцентного контура при реализации на материале словоформ, содержащих и не содержащих заударные слоги, например, *Вася* и *Джон*, показывает, что финальный – нисходящий – компонент тональной структуры фиксируется на последнем заударном слоге, как в *Вася-я-я* (ср. также *Иван Федорович*), и смещается на ударный слог, как в *Джо-о-он*, если заударных нет. Тогда ударный гласный приобретает двухчастный характер, между двумя компонентами которого возникает гортанная смычка. Тем самым фактически создается дополнительный слог. Если обращение состоит более чем из одной словоформы, носителем акцента служит конечная: a. *Гвардии* ***сержа-ант***!; b. *Сержант* ***гваардии-и***!; c. *Гости* ***дорогие-е***!; d. *Дорогие* ***гости-и***. Минимальные пары a vs. b и c vs. d показывают, что акцентоноситель здесь выбирается в соответствии с принципом линейной структуры, т.е. синтаксическая структура на выбор акцентоносителя не влияет. Таким образом, русский язык отступает здесь от своего базового – синтаксического – принципа выбора акцентоносителя в коммуникативном компоненте речевого акта, основанного на синтаксических приоритетах. О синтаксических приоритетах при выборе акцентоносителя см. Янко 2001: 68-83.

Другой вокативный тип представлен в ситуации зова адресата, когда слушающий не знает, находится ли адресат в пределах слышимости. Говорящий обращается с призывом к невидимому адресату в пустой или в темной комнате, в прихожей дома, когда неизвестно, на месте ли хозяева, к спящему, не будучи уверенным в том, что спящий его слышит, говорящий зовет слушающего в толпе, в лесу среди деревьев. Рассмотрим тонограмму.



Хозяйка-а?! *Маруся?! Где это ты та-ам?! Хеллоу-?!* *Есть кто-нибудь?!* *Иван Иваны-ыч?!*

092-Yanko_T:_Layout 1 18.05.2008 16:57 Page 598

Здесь мы наблюдаем низкий ровный тон на предударных, за которым следует подъем на одном из ударных слогов[4]. Затем возникает падение тона на заударных, если они есть. Завершается контур растянутым, ровным и достаточно высоким тоном. Если заударного слога нет, он создается путем растяжения конечного слога и расщепления его на два фрагмента с помощью смычки так же, как при зове адресата, значительно удаленного от слушающего. Акцентный пик в двух рассмотренных здесь апеллятивных контурах приходится на последний заударный слог конечной словоформы акцентоносителя, если обращение состоит из именной группы: *Иван* ***Иваны-ы-ыч***; *Господин* ***профессо-ор***. В ориентации акцентного пика на заударный слог с искусственным созданием подобия такого слога в словоформах, где заударных слогов нет, мы видим высокую степень идиоматичности рассматриваемых типов речевых актов: стандартные стратегии строятся с ориентацией, наоборот, на ударные слоги.

2.1. Другие уникальные свойства обращений

Просодия – это не единственная уникальная характеристика обращений. О нестандартных свойствах обращений существует большая литература. Интересной особенностью т.н. нового русского вокатива служит противоречащая русской фонетике способность последнего согласного не оглушаться: *Федь*! [f'ed'] (Панов 1997: 108-110). Другая особенность, которую выделяют авторы Yadroff 1996, Corbett 2007, Daniel, Spencer (in press), Даниэль (в печати), это отсутствие беглой гласной в формах типа *Мишк*!, ср. нулевую форму генитива множественного числа с прояснением беглой гласной *Мишек*. Мы, впрочем, не видим в формах *Мишк*! и *Людк*, *а Людк*! ничего выходящего за пределы естественного развития языка: новый вокатив образовался уже после того, как редуцированные пали. Еще один параметр, имеющий непосредственное отношение к составу словаря, это способность лексем употребляться в функции обращения. Так, А. Цвикки на материале английского языка сделал заключение, верное и для русского, что слова *доктор* и *врач* (и их английские эквиваленты) различаются по способности функционировать в функции обращения (Zwicky 1974).

Ниже мы рассмотрим две особенности, которые отличают обращения от других типов речевых актов: сообщений, вопросов и императивов. Это 1) особые принципы выбора акцентоносителя и 2) уникальное – требующее словарной фиксации – поведение в составе обращения местоимений *мой*, *твой*, *наш* и *ваш*. Положения 1) и 2) связаны таким образом, что особенности местоимений проявляются именно при выборе акцентоносителя.

Обратимся к проблеме выбора акцентоносителя. Вначале определим, каков стандартный – «дефолтный» – принцип выбора акцентоносителя в именных группах, не отягощенных контекстами обращения, мольбы, упрека, угрозы, недоумения, контраста, эмфазы. В именной группе с адъективным определением, которое не имеет зависимых слов и расположено перед определяемым словом, ударно определяемое имя – *моя* ***дочь*** (*Это моя* ***дочь***), *ваша* ***честь*** (*На карту поставлена ваша* ***честь***), *твоя* ***милость*** (*Велика твоя* ***милость***), *наш* ***друг*** (*Ему помог наш* ***друг***), см. Ковтунова 1976: 65-67; Светозарова 1993. В группах с несогласованным определением ударно определение, которое в норме расположено после определяемого слова: *начальник* ***смены***. В именных группах типа «имя-фамилия» ударна фамилия, которая расположена после имени: *Вася* ***Иванов*** (*Меня зовут Вася* ***Иванов***). В именах-отчествах ударно отчество (*Иван* ***Иванович***), в конструкциях «титул-имя» ударно имя (*профессор* ***Иванов***). Такой принцип выбора акцентоносителя мы называем синтаксическим.

В обращениях (но не только в них) синтаксический принцип выбора акцентоносителя действует не всегда. Выбор акцентоносителя в обращении определяется разнообразными прагматическими факторами с образованием целого спектра вокативных иллокутивных сил, таких как попытка говорящего докричаться до слушающего (*Ва-а-с-я-я*!), стремление упрекнуть (*Ива-ан Иваныч*, *ну как же так*?), остановить уходящего или отвернувшегося (*Иван Ива-аныч*, *ну куда же вы*?). Эти иллокутивные силы различаются характерной для каждого из них мелодикой и принципом выбора носителей акцентных пиков. В выборе акцентоносителей можно наблюдать следующие закономерности. Если слушающий находится близко от говорящего и обращение включает более одной словоформы, ударна первая словоформа (***Марья*** *Ивановна*, *пойдемте обедать*; ***Молодой*** *человек*, *купите букетик*; ***Начальник*** *смены*, *пройдите к пульту*), а если далеко – иконически ударна последняя словоформа (*Марья* ***Ивановн-аа***! *Идите к на-ам*!). Другой фактор – психологическая дистанция между говорящим и слушающим. При неофициальном тоне ударна первая словоформа (***Иван*** *Петрович*, *что за радость*!), при официальном – последняя (*Глубокоуважаемый Иван* ***Петрович*****!** *Примите наши поздравления*). Пространственная и психологическая дистанция сочетаются в одном обращении таким образом, что побеждает стратегия того параметра, который имеет значение удаленного общения, – пространственного или психологического: если говорящий и слушающий находятся близко друг от друга, но общаются в официальном тоне, используется стратегия, маркирующая психологическую дистанцию. Анализ этих стратегий говорит о том, что в обращениях синтаксиси-

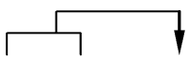
[4] Ответ на вопрос о том, как выбираются акцентоносители падения и подъема, мы здесь опускаем.

092-Yanko_T:_Layout 1 18.05.2008 16:57 Page 599

ческий принцип может уступать место линейному. Это нетривиальный факт русского языка.

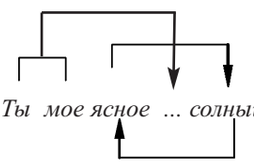
Линейный принцип – не единственный принцип выбора акцентоносителя в обращениях. Рассмотрим обращения в условиях, когда говорящий преследует цель поддержания контакта со слушающим и необходимость инициации общения отсутствует: *Подвинься*, *Зин*; *Я*, *Вань*, *такую же хочу*; *Вы*, *Марь Ванна*, *не волнуйтесь так*. Проблемы выбора акцентоносителя здесь нет, потому что имя слушающего вообще лишено каких бы то ни было релевантных просодических признаков, т.е. с точки зрения коммуникативно релевантных акцентов «безударно», и имеет тенденцию занимать второе место в предложении; об обращениях в позиции Ваккернагеля см. Renou 1936: 61; Gonda 1971: 146-147. Этот принцип мы называем безакцентным.

Еще один принцип выбора акцентоносителя реализуется с двумя акцентными пиками – в начале и в конце обращения, т.е. с двумя акцентоносителями. Данный принцип действует в обращениях, по иллокутивной силе близких восклицаниям. Такую модель Брызгунова называет интонационной конструкцией ИК-5 (Русская грамматика 1982: 107): ***Дорогой*** *Иван* ***Петрович***! с подъемом на *дорогой* и падением на *Петрович*; ***Горе*** *ты мое* ***луковое*** с подъемом на *горе* и падением на *луковое*, ***Ясное*** *ты мое* ***солнышко***! с подъемом на *ясное* и падением на *солнышко*. Для обращений данного типа характерна дислокация[5] – расщепление именной группы с помещением других словоформ, если они есть, в образовавшуюся при разрыве нишу. Так, при дислокации в исходной структуре *Ты мое ясное солнышко*! фрагмент *ясное солнышко* разрывается, словоформа *ясное* помещается в начало предложения, *солнышко* – в исход, а словоформы *ты* и *мое* располагаются в разрыве между *солнышко* и *ясное*:



a. Ты мое ясное ... солнышко → Ясное ты мое солнышко!

Кроме дислокации, данный тип может сопровождать и инверсия, когда *ясное* и *солнышко* меняются местами:



b. Ты мое ясное ... солнышко → Солнышко ты мое ясное!

Для полноты картины отметим, что синтаксический принцип выбора акцентоносителя также не чужд обращениям, в частности, так называемым обращениям-характеризациям (Пешковский 1956: 407, Ковтунова 1986: 105), лишенным собственно апеллятивной функции. Этот тип широко используется в поэтической речи: *Волшбного шептанья полный* ***лес*** ... (Баратынский); *Вы*, *чьи широкие шинели напоминали* ***паруса***... (Цветаева)[6]. В развернутых обращениях-характеризациях акцентоноситель выбирается в соответствии со стандартным принципом русского языка. В разговорной речи обращения-характеризации также широко используются, особенно при восторженно-ласковых обращениях и в ругательствах: *Машенька*! *Красавица*!; *Брось прикидываться, медвежья башка* (Коваль).

Итак, обращения обнаруживают четыре стратегии выбора акцентоносителей: 1) синтаксическую (как в развернутых обращениях-характеристиках), 2) линейную (как в ***Иван*** *Петрович*! и *Иван* ***Петрови-ич***!), 3) безакцентную (как в *подвинься, Зин*) и 4) двухакцентную (как в ***Дорогая Маша***; ***Неподкупный*** *вы* ***наш***).

В следующем подразделе будет показано, что особенности местоимений в структуре обращений требуют словарной фиксации.

3. Мой, ваш, наш и твой в составе обращений

Начнем с местоимения *ваш*. В современных русских обращених оно встречается весьма ограниченно: в обращениях к судье (*Ваша честь*!), к патриарху (*Ваше святейшество*!), в историзмах (*Ваше благородие*) и в переводах с других языков (*Ваше высокопреосвященство*). Оно способно реализовать следующие принципы

[5] Термин «дислокация» введен И.И. Ковтуновой (1976: 120-121).

[6] Примеры заимствованы из книги Ковтунова 1986: 96-127.

092-Yanko_T:_Layout 1 18.05.2008 16:57 Page 600

выбора акцентносителя: линейный (***Ваша*** *честь*, *у защиты вопрос к свидетелю*), безакцентный (*В этом*, *ваша честь*, *моему доверителю было отказано*) и двухакцентный (***Ваша честь***!).

Местоимение *мой* входит в большое количество закрепленных в русском узусе обращений: *мой дорогой*, *милая моя*, *мой повелитель*, *моя радость*, *счастье мое, солнышко мое*, *горе мое луковое*, *сын мой*, *рыбка моя*, *птичка моя*. Выше было показано, что при линейном принципе ближней апелляции к слушающему, носителем акцентного пика становится начальная словоформа, которой, в частности, может быть и притяжательное местоимение: ***Ваша*** *честь*!. У местоимения *мой* способность быть акцентоносителем в таких обращениях иная, чем у *ваш*. В составе галлицизмов *мой господин*, *мой император*, *мой капитан* и *мой генерал* местоимение *мой* может выступать в роли акцентоносителя всего словосочетания: ***Мой*** *господин*, *пожалуйте к столу.* Атоническая форма здесь также возможна и даже более вероятна: *Мой* ***господин***, *пожалуйте к столу.* В составе устойчивых обращений *мой дорогой*, *моя милая* перенос акцента на *мой* возможен только в речевых актах упрека, сетования и уговора – контекстов, которые стимулируют перенос акцента на первый компонент словосочетания: ***Мой*** *дорогой*, *ну как же так*?; ***Дорогой*** *мой*, *ну как же так*? В отсутствие упрека переноса акцента на *мой* не происходит: *Мой* ***дорогой***, *пойдем позавтракаем* (ср. ?***Мой*** *дорогой*, *пойдем позавтракаем*). В других контекстах местоимение *мой* сохраняет атоническую форму даже в контексте упрека: ***Солнышко*** *мое*, *ну как же так*?! (****Мое*** *солнышко*, *ну как же так*?); *Моя* ***радость***, *ну как же так*?! (****Моя*** *радость*, *ну как же так*?!); ***Доченька*** *моя*, *ну как же так*?!. Регулярному акцентированию местоимение *мой* подвергается только в исходе рамочных структур с дислокацией и инверсией, как в ***Миленький*** *ты* ***мой***, где на *миленький* фиксируется подъем, а на *мой* – падение.

Итак, в отличие от *ваш* местоимение *мой* имеет сильную тенденцию к безакцентному употреблению. Это словарное свойство местоимения *мой*. Объяснением этому может служить высокая частотность местоимения *мой* в обращениях. Аналогичный механизм употребления и преобразования староиспанского обращения *Vuestra Merced* ‘Ваша милость’, давшего современное *usted* ‘вежл. Вы’, Вяч. Вс. Иванов объясняет «статистически обусловленным изнашиванием самых употребительных слов языка, которое осуществлялось на больших отрезках времени» (Иванов 2004), ср. также русское *сударь*, за короткое время превратившееся в т.н. словоер: *да-с*.

Местоимение *наш* в обращениях представлено в двухакцентных конструкциях (***Дорогой*** *ты* ***наш***) и безакцентно (*Наш* ***дорогой*** *Иван* ***Петрович***!). В контексте линейного акцентирования при ближней апелляции по нашим наблюдениям не встречается: ****Наш*** *дорогой*! ****Наша*** *Маша*.

И наконец, местоимение *твой* в формировании русских обращений фактически не участвует. В письменной базе данных оно встретилось в обращении-историзме *твоя милость* безакцентно только один раз: *Рассуди же здраво, твоя милость, чего ты хочешь от меня*?

***

На примере речевого акта обращения была рассмотрена проблема представления просодической информации в словаре языка.

Список литературы

1. Апресян Ю.Д. Прагматическая информация для толкового словаря // Логический анализ языка. Прагматика и проблемы интенсиональности. М.: Институт языкознания, 1988.

2. Богуславский И.М. Сфера действия лексических единиц. М.: Языки русской культуры, 1996.

3. Булыгина Т.В., Шмелев А.Д. Языковая концептуализация мира (на материале русской грамматики). М.: Языки русской культуры, 1997.

4. Даниэль М.А. (в печати) Звательность как дискурсивная категория. Несколько гипотез.

5. Иванов Вяч. Вс. Лингвистика третьего тысячелетия: вопросы к будущему. М.: Языки русской культуры, 2004.

6. Ковтунова И.И. Современный русский язык. Порядок слов и актуальное членение предложения. М.: Просвещение, 1976.

7. Ковтунова И.И. Поэтический синтаксис. М.: Наука, 1986.

8. Падучева Е.В. О семантике синтаксиса. М.: Наука, 1974.

9. Падучева Е.В. Давно и долго // Логический анализ языка. Язык и время. М. 1997.

10. Панков Ф.И. Наречная темпоральность и ее речевые реализации. АКД. М., 1996.

11. Панков Ф.И. Синтаксические позиции русских наречий. Позиции адвербиальных синтаксем // Слово. Грамматика. Речь. Вып. VII: Сборник научно-методических статей по преподаванию РКИ. М., 2005.

092-Yanko_T:_Layout 1 18.05.2008 16:57 Page 601

12. Панков Ф.И. Синтаксические позиции русских наречий. «Членопредложенческий ранг» адвербиальных словоформ // Слово. Грамматика. Речь. Вып. VIII: Сборник научно-методических статей по преподаванию РКИ. М., 2006.

13. Панов М.В. Современный русский язык. Фонетика. М. 1997.

14. Пешковский А.М. Русский синтаксис в научном освещении. Изд. 7. М.: Просвещение, 1956.

15. Русская грамматика Т. 1, М.: Наука, 1982.

16. Светозарова Н.Д.Акцентно-ритмические инновации в русской спонтанной речи // Проблемы фонетики. 1993. Вып. I. М.

17. Янко Т.Е. Коммуникативные стратегии русской речи. М.: Языки русской культуры, 2001.

18. Corbett G. Determining morphosyntactic feature values: the case of case // Case and grammatical relations. Papers in honour of Bernard Comrie. Oxford: Oxford University Press, 2007.

19. Daniel M., Spencer A. (in press). Vocative: an outlier case.

20. Gonda J.Die Indischen Sprachen: Erster Abschnitt: Old Indian. Leyden; Cologne, 1971.

21. Pierrehumbert J.B. The Phonology and Phonetics of English Intonation. PhD thesis. MIT. 1980.

22. Renou L. Etudes de grammaire sanscrite. Paris, 1936.

23. Yadroff M. Modern Russian Vocatives: a Case of Subtractive Morphology // Journal of Slavic Linguistics. 1996. Vol. 4.

24. Yanko T.E. The Communicative Effects on the Interaction between the Verbal Aspectual Categories and Temporal Adverbials in Russian // Journal of Slavic Linguistics. 2003. Vol. 11. N 1.

25. Zwicky A. Hey, whatsyourname! // Papers from the Tenth Regional Meeting of the Chicago Linguistics Society. Chicago: Chicago Linguistics Society, 1974.

093-Oja_A:_Layout 1 18.05.2008 16:31 Page 602

ВЫБОР ЯЗЫКА В ИНТЕРНЕТ-КОММУНИКАЦИИ МЕЖДУ РУССКИМИ И ЭСТОНЦАМИ

CHOOSING LANGUAGE IN INTERNET CONVERSATION BETWEEN RUSSIANS AND ESTONIANS*

Anni Oja *(anni.oja@tlu.ee)*
Tallinn University

В статье рассматривается межъязыковая коммуникация на эстонском сетевом портале rate.ee. Рассмотрены диалоговые акты между эстонцами и русскими (комментарии к фотографиям) для того, чтобы определить факторы выбора языка. Большинство этих факторов относится к ситуации: идентичности участников, удобству и цели общения), но некоторые стратегии применяются подсознательно и исходят из неписанных правил вежливости и выбора нейтрального языка с использованием подходящей лексики.

1. Introduction. Online communities as communities of practice: learning through communication

In recent years the Internet has became significant „meeting place“ for all kind of languages, social groups, nations. Being founded in USA with major language English it has succeeded in world-wide technological revolution and given new meaning to the phrase „intercultural communication“. Statistics [1] shows extremely rapid growth of Internet usage in non-English regions, for example in Middle East in period 2000-2007 the growth has been 920.2 %.

Since the very beginning of Internet access virtual communities have existed. The technical base of community has changed through the years (bulletin boards, telnet chatrooms, multi-user dungeons, mailing lists, forums, massive multiplayer online role-playing games, social networking sites etc), but the living force of communities appearing in all kinds of environments shows the growing demand of communication and belonging between Internet users.

Virtual communities generally act as the communities of practice. Communities of practice, concept worked out by Etienne Wenger, are „groups of people who share a concern or a passion for something they do and learn how to do it better as they interact regularly“ [2]. This may be not intentional, but appear as a coeffect of other communication purposes. In online dating environments these purposes may be quite personal at the first sight – finding suitable partner, promoting yourself, building image of success. Still there is tendency for evolving communities with some collective identity and non-written communication rules. By Wenger [3] the communities of practice have produced a shared repertoire of communal resources – language, routines, sensibilites, artifacts, tools, stories, styles, and so forth. To be competent is to have access to this repertoire and to be able to use it appropriately. So in such environments language use is the competence to be learned and developed. Textual communication is pivotal to understanding virtual communities, for it is the means for participants of those virtual environments to create, affirm, or change shared meaning and culture [4].

2. Material and method. Interculturality in Estonian portal rate.ee

Community of practice to be analysed is the user base of the most popular Estonian communication website rate.ee, which was started in 2002 as a picture rating portal (comparable to Hotornot.com), but nowadays acts as a multifunctional full-featured web environment with options of weblog, gallery, chatroom, gaming corner and news feed. It's interface is only Estonian, so it is mainly used by inhabitants of Estonia who have gained at least some knowledge of Estonian language. Though rate.ee's userbase is Estonian-centered, the community is still multilingual. During the last census in Estonia [5] the native language was Estonian for 921,817 inhabitants of the country, Russian for 406,755 inhabitants, Ukrainian for 12,299 inhabitants, Belorussian for 5,197 inhabitants, Finnish for 4,932 inhabitants and Latvian for 1,389 inhabitants. It is difficult to say exact amount or percentage of Russians in rate.ee as there is no field for nationality, but language skills give us some sense of it. In statistics of rate.ee [6] 77 % of users have declared their Estonian skill as „very good“ and 20 % of users give the same rating to their Russian skill. In general 90 % of users have at least some knowledge of Estonian and 67 % have at least some knowledge of Russian.

* This work is supported by the ETF grant no 6147.

093-Oja_A:_Layout 1 18.05.2008 16:31 Page 603

Author has participated in rate.ee community since its beginning in 2002, since 2006 in role of moderator (power-user with moderating abilities). Moderator has better overview of userbase and communication trends, accepting or rejecting new pictures, helping users with their problems and removing users with incorrect data.

Texts to be analysed are picture comments (public text) from rate.ee. Often the picture viewer does not check the profile of picture owner, so he/she gets the information about ethnicity/language skills in other ways: from language used in picture signature, from comments of other commentators and picture owner's responses, sometimes from username or picture context.

The selection of language material is made of 1) comments posted by probable Russians to probable Estonians, 2) comments posted by probable Estonians to probable Russians and 3) comments posted by probable Russians to probable Russians, but in Estonian. The determiners of nationality in this context are language skills defined by users themselves (for example „Russian – very good, English – intermediate, Estonian – poor"), names (generally Estonian and Russian names differentiate clearly), school information (Estonian or Russian school), language used to describe him/herself, language used in weblog, language used for commenting others. For analysis were chosen 150 user accounts that fitted restrictions mentioned before.

The comments will be analysed and compared to see the general and special features in different language situations. This helps to define environmental language details that may have been learned from each other in interlingual communication and the ones that are universal or depending more on situation than environment.

Analysis is limited to available information: it is possible that some users play with identity or that their real-life identity is multilingual (for example in bilingual and -national families), the language proficiency is also defined by user, so there exists a chance that it is overrated or underestimated. Anyway, we can only relay on available data, considering the chance that background information may be somewhat invalid or inprecise. Still the information is quite valid as the community regulates itself. Estonia is small country and people know each other quite well, so when somebody defaces his/her data, quite soon somebody notices it and informs the moderators of the site. Being in the role of moderator, the author has received many such complaints and always asked for reason to complain. More common answers have been of type „I would like to be moderator, too" (to get the virtual honour and power), „I have correct data myself and I want other users to be honest, too" and „he/she is lying, there is no such person in our class/village/etc, I couldn't stand it". When moderators get information about fake account or fake data, the account will be closed. It means that the owner of account loses all his/her messages, contacts, virtual reputation, membership of private clubs and social network. He/she can create new account, but it takes time to restore all his/her virtual „life". There is one more reason to present correct data: people looking for real-life relations want to meet others face-to-face and for that they have to present „real self" to avoid embarrassing meetings. This topic has been discussed in rate.ee forum and youngsters tell, that they better present their (almost) real presence in rate.ee to get real friends.

3. Analysis

The conditions of choosing one or another language were examined. In most cases, when commentator and the picture's owner were not acquainted, the commentator tried to use the language of preliminary comments.

(1) ilus oled :) [you are pretty]
from: male, 21, Russian very good, Estonian: intermediate, English: intermediate
to: female, 16, Russian: very good, Estonian: intermediate, English: intermediate
Previous comments, written by Estonians in Estonian: *kaunitar!!! (k) [beauty!!!]*
täiega armas [absolutely cute]

If he/she was not familiar with that language, then mainly English was used. While responding to comment, the picture's owner tried to use same language or when he/she was not familiar with that language, he/she used English or responded with emoticon or some very common short neutral and polite word (for example *tnx*, shortened version of combination of Estonian *tänan* (meaning *thanks*) and English *thanks*.

(2) Ouuuch, My godnesssss....you are soo Pretty!!!!)
response*: no, you are!*
from: female, 16, Russian: very good, Estonian: very good
to: male, 15, language skills undefined, but studying in Estonian school and having Estonian name

Sometimes the disorientation appeared, when comment included more than one language.

(3) t6 samaja samaja ,,iludus-. [iludus = beauty]
response: *Tänan sind;)* [Thank you]
from: male, 16, Russian: very good, Estonian: poor
to: female, 16, language skills undefined, but commenting others only in Russian)

093-Oja_A:_Layout 1 18.05.2008 16:31 Page 604

There was also one interesting correction: commentator used firstly Estonian, then probably noted other comments in Russian and after one minute wrote second comment in Russian with generally same meaning.

(4) oilt on paris seksikas.kuid voiks seal olla rohkem sartsu. [the picture is kind of sexy, but there might be more electricity]

After one minute: *fotka seks.no tam egoto ne hvataet.tebe nado bolo bo bolse fantazii*

from: male, 18, language skills: Russian:very good, Estonian: very good

to: female, 16, language skills: Russian: very good, Estonian: poor

In some cases the first comment was in non-native language of picture's owner, then response was very neutral (emoticon or just one word) and the next comment was already given in native language. It is possible, that neutral response indicates communication breakdown, gives a signal to change the communication code.

Some users responded only to comments produced in their native language, ignoring foreign-language comments. This silence normally ends the conversation, as the first commentator does not get any encouragement. Neutral answer, on the contrary, still leaves the possibility of following discussion, even when having no actual or positive meaning, whereas no answer is understood as negative, cancelling signal.

Commenting in non-native language tends to be shorter, more neutral and more unoriginal than commenting in native language. Such comments are typically *7p* (seven points), *10p* (ten points), *super*, *wau*, *kena* (pretty), *lahe pilt* (cool picture), *klassnaja*. English words and phrases are very common for both Estonians and Russians: *nice, beautiful, best, sexy, sweet, cute, very cool, UR2GOOD, nice pic*. Such comments will get no or minimal polite response.

The length and contents of comment are also related to purpose of communication. In social network portals one quite widespread aim is to find partner. In rate.ee this is specially supported in environment as every userpage has a relationship status field and extra field for chosen one. One can set another user as his „chosen one", and when that user sets him, too, they are displayed on each other's pages. Having accepted „chosen one" is so important, that even some 7-year-olds have this slot filled.

For example test query of comments of *male users in age 15-22 without girlfriend, commenting female users with opposite language proficiency* gave more long comments than *female users in age 15-22 with boyfriend, commenting male users with opposite language proficiency*. When communication had a purpose, the motivation was recognisably high: dialogue was long, there were many mini-dialogues between users A and B and they managed to find a way to communicate – writing with mistakes, making copy-paste of other comments, combining languages:

(5) oled vaga kaunitar(KL)= [you are very beauty, in correct Estonian it should be *oled väga kaunis*]

from: male, 18, Russian: very good, Estonian: very good

to: female, 18, with Estonian name, studying in Estonian school

Once Estonian language was clearly used in response to give a sign of unwanted communication, to cancel communication in „polite way":

(6) ja tebe eshe i sovetov to ne daval :))))) gde ti moj sovet uvidela to?eto rekomendacii bili)) ja ne kazdomu 4eloveku sovet daju ;)relax,it's not so deep)))))))

response: *Aitab spämmida siin :D* [incorrect Estonian: Enough of spamming here, in correct Estonian it should be *Aitab siin spämmimisest*]

from: male, 21, Russian: very good, Estonian: intermediate

to: female, 24, Russian: very good, Estonian: very good, usually writing in Russian

4. Discussion

Background of language-political situation was intentionally ignored here as there are no sufficiently reliable statistical researches about cross-lingual and cross-cultural attitudes in Estonia. Still in analysed material there were no examples of aggressive language attitudes like „speak my language!" or „I know your language, but I don't want to speak it". There is a chance that users, who never responded to comments in another language, may „protest" in this way, but based on their profiles it is more likely that they just did not understand the text and were not able to compose answer. The indicators to assume it are 1) user's language skills, for example Estonian is defined as „poor", all information is presented in Russian and user always comments others only in Russian, 2) living area and school, in some regions the level of teaching Estonian is still primitive and students are not able to get proper Estonian skills. Interviews show that young Russians in Estonia are generally interested in learning Estonian, but for improving it they need Estonian-speaking environment [7].

Conversation situation affects the language choice a lot:

1. what language is used in other comments?
2. do the commentator know the person he/she comments?

093-Oja_A:_Layout 1 18.05.2008 16:31 Page 605

3. what is the purpose of text? Do the commentators hold long conversation and for example aim to get to know each other, presenting themselves and proposing to meet in real life?

4. is it comfortable to use this language: for commentator, for a person commented? Does the commentator want to create comfortable or uncomfortable situation?

Environmental learning helps users to understand the inside rules of rate.ee language politeness:

1) which vocabulary is accepted as neutral and neutral-positive?

2) what kind of response is suitable as „polite enough"?

3) how to cancel unwanted conversation in polite way, still making the cancellation clear?

In future it is possible to go more deep, creating subcorpus of interlingual communication in rate.ee and including other nations and languages. Estonian-Russian code-switching has been often viewed from the (Estonian) language teacher's point of view (for example studies based on Estonian Interlanguage Corpus [9]). Still some studies view Russian's impact on Estonian language, for example Tene Üprus has measured Russian accent in Estonian spontaneous speech and discovered that Estonians may adapt „Russian accent" to communicate more efficiently with Russians [8]. Written Internet communication does not give us much information about accent or learning, but at the same time helps to shine light on general language trends in normal spontaneous conversations, providing huge text collection composed by language users themselves.

References

1. Internet Usage Statistics 2000-2007. http://www.internetworlstats.com/stats.htm (last retrieved 18.02.2008)

2. Wenger, Etienne. Communities of practice, a brief introduction. http://www.ewenger.com/theory/ (last retrieved 18.02.2008)

3. Wenger, Etienne. Communities of Practice and Social Learning Systems. // Knowing in Organizations: A Practice-Based Approach. (Eds) Nicolini, Davide; Gherardi, Silvia; Yanow, Dvora. M. E. Sharpe, 2003. p. 80

4. Burnett, G., Chudoba, K. M., Dickey, M. H., Kazmer, M. M. Inscription and interpretation of text: a cultural hermeneutic examination of virtual community. // Information Research (online journal), Vol. 9, 4, paper 162, 2003. http://informationr.net/ir/9-1/paper162.html (last retrieved 18.02.2008)

5. Statistics Estonia. Populational Census 2000. http://pub.stat.ee/px-web.2001/I_Databas/Population_Census/Population_Census.asp (last retrieved 20.02.2008)

6. Rate.ee statistics. http://www.rate.ee/reports.php (last retrieved 20.02.2008)

7. Küün, Elvira. Mitte-eestlastest noorte kohanemine eestikeelses töökeskkonnas. // EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 2. Estonian Papers in Applied Linguistics 2. Editors: Helle Metslang (Helsinki/Tallinn), Margit Langemets (Tallinn). Associate editor: Maria-Maren Sepper. Tallinn: Eesti Keele Sihtasutus, 2006.

8. Üprus, Tene. Eestlaste keeleline kohanemine eesti-vene segarühma suhtluses. // Keel ja Kirjandus, Vol 5, pp. 379-388, 2006.

9. Eslon, Pille (editor). Tallinna Ülikooli keelekorpuste optimaalsus, töötlemine ja kasutamine. Eesti filoloogia osakonna toimetised 9. Tallinna Ülikooli Kirjastus 2007.

094-Partee_B:_Layout 1 18.05.2008 16:32 Page 606

СИММЕТРИЯ И СИММЕТРИЧНЫЕ ПРЕДИКАТЫ
SYMMETRY AND SYMMETRICAL PREDICATES*

Partee Barbara *(partee@linguist.umass.edu)*
University of Massachusetts, Amherst, MA, USA;
Russian State University for the Humanities

Цель этой статьи – проанализировать разницу между математическими определениями симметрии и понятием симметрии, которое бы наилучшим образом соответствовало лингвистическим обобщениям. Это требует тщательного анализа лингвистического «поведения» симметричных и несимметричных предикатов.

1. Background

"Symmetrical predicates" have distinctive linguistic properties in many languages. But the concept of "symmetry" merits its closer examination, especially in the light of the controversial claim by the psychologist Amos Tversky [1] that the concept 'similar', a standard example of a symmetrical predicate, is in fact not symmetrical. Tversky's evidence includes the fact that experimental subjects generally rate (1a) as holding to a higher degree than (1b).

(1) a. *North Korea is similar to Red China.*
b. *Red China is similar to North Korea.*

Lila Gleitman and colleagues argue in an interesting paper [2] that 'similar' is symmetrical, and that the difference in judgments reflects the independent contribution of figure-ground differences encoded in the syntax. They argue in support of a robust linguistic distinction between symmetrical and "asymmetrical" predicates. Gleitman *et al* use a semantic paraphrase test as a central property in characterizing linguistically symmetrical predicates in English: does the *intransitive* version of a given predicate have a meaning close to the meaning of an overt reciprocal with the corresponding *transitive* version? This test is illustrated in (2) and (3) below, where (2a) and (2b), with symmetrical *meet*, are close in meaning, but (3a) and (3b), with the "asymmetrical" *drown*, are not.

(2) a. *John and Bill meet.*
b. *John and Bill meet each other.*

(3) a. *John and Bill drown.*
b. *John and Bill drown each other.*

Gleitman *et al*'s paper analyzes symmetrical and what I will call "quasi-symmetrical" or "sometimes-symmetrical" predicates in English, including verbs (*meet*, *kiss*), and adjectives (*similar*), to which I will add nouns (*sibling, brother*). Their paper addresses and solves the mysteries raised by Tversky's work concerning the apparent non-symmetrical behavior of symmetrical predicates like *similar*.

The arguments in the paper are convincing; at the same time, Gleitman *et al*'s uses of the terms "symmetrical" and "asymmetrical" do not always fit the standard mathematical definitions, given in (4). (Variation in definitions is discussed in Section 2.)

(4) a. A relation R is symmetrical iff for all x, y: if R(x,y), then R(y,x).
b. A relation R is asymmetrical iff for all x, y: if R(x,y), then ¬ R(y,x).
c. A relation R is non-symmetrical iff if is not symmetrical.

One goal of this paper is to analyze the differences between the standard mathematical definitions of symmetry and a concept of symmetry that would fit best with observed linguistic generalizations. In order to try to modify the mathematical definitions to better fit the linguistic facts, we have to look at the linguistic facts more carefully as well.

* This material is based upon work supported in part by the National Science Foundation under Grant No. BCS-0418311 to Barbara H. Partee and Vladimir Borschev. For discussion and suggestions, I am grateful to Lila Gleitman over several years, and to Muffy Siegel, Tony Kroch, and other members of the audience at the University of Pennsylvania where a preliminary version of this work was presented in October 2007.

094-Partee_B:_Layout 1 18.05.2008 16:32 Page 607

2. Definitions and examples

2.1. The standard terminology and its consequences, with examples

The definitions of *symmetrical, asymmetrical, non-symmetrical* in (4) above are taken from Partee, ter Meulen, and Wall [3] (PtMW) (where the equivalent *symmetric, asymmetric, non-symmetric* are used), which in turn followed such classics as [4]. There is a fourth term in this family, *antisymmetrical*, defined in [3] and elsewhere as follows:

(5) A relation R is *antisymmetrical* iff for all x, y: if R(x,y) and R(y,x), then x = y.

Typical examples: ≤ is antisymmetrical, whereas < is asymmetrical. An *antisymmetrical* relation is asymmetrical except for possible instances of *x*R*x*. Turning it around, an *asymmetrical* relation is one that is antisymmetrical and irreflexive.

We begin by reviewing some consequences of the standard definitions and some examples.

First of all: Are the four properties mutually exclusive? No.

(i) Every asymmetrical relation is non-symmetrical. One can call *asymmetrical* a 'strong' negation of *symmetrical*: it means "never symmetrical". *Non-symmetrical* just says "sometimes not": R is non-symmetrical iff it is not symmetrical.

(ii) Every asymmetrical relation is also antisymmetrical. If it's asymmetrical, the if-clause in the definition of *antisymmetrical* is never satisfied, so the entire statement is vacuously satisfied.

(iii) Can a symmetrical relation ever have any of the *a-*, *non-*, *anti-* symmetry properties? Yes: the empty relation, for instance, is symmetrical, asymmetrical, and antisymmetrical. And the identity relation on some domain D, which has all and only pairs of the form <a,a> for a in D, is both symmetrical and antisymmetrical.

Secondly, are the four properties exhaustive? That is, does every relation have at least one of those properties? Yes, because *symmetrical* and *non-symmetrical* are complements: every relation on a given domain A × B must be either symmetrical or non-symmetrical.

The relation *father of* is asymmetrical, since for all x, y, if x is the father of y, then y is not the father of x.

The relation *sibling of* (i.e. *brother or sister of*) is symmetrical, since for all x, y, if x is a sibling of y, then y is a sibling of x.

What about the relation ***brother of****:* is it true that for all x,y, if x is the brother of y, then y is the brother of x? This can't be answered without specifying the **domain** of the relation.

(a) On the domain of all humans, *brother of* is neither symmetrical nor asymmetrical, but non-symmetrical: if x is brother of y, y may be brother of x or sister of x.

(b) On the domain of all male humans, *brother of* is symmetrical.

We'll return to this case in Section 4.2, because *brother* behaves linguistically like a symmetrical predicate in sentences like (6).

(6) *John and Bill are brothers.*

2.2. Colloquial terminology

The terminology: When talking about the *brother-of* relation, students often say "sometimes it's symmetrical and sometimes it's asymmetrical." Or "it has both symmetrical and asymmetrical instances." But according to the definitions, *symmetrical* and *asymmetrical* are properties of *relations*, and it doesn't make sense to apply those terms to *instances*, or to say that a relation has such a property "sometimes". But it's clear what the students mean, and we can give new definitions to extend the terminology in these ways.

Defining "symmetrical instances":

(7) (a) "Sometimes *brother* is symmetrical" and "*Brother* has symmetrical instances" can both be defined as follows: there are pairs a,b such that a is brother of b and b is brother of a.

(b) *Brother* has asymmetrical instances" can be similarly defined as saying that there are pairs a,b such that a is brother of b and b is not brother of a.

Note that in both cases we start from an instance of aRb, and then ask whether bRa holds. Since either it does

094-Partee_B:_Layout 1 18.05.2008 16:32 Page 608

or it doesn't, there are only two cases, symmetric or asymmetric, for 'instances'.

Given this notion of symmetrical and asymmetrical *instances*, the original concept of a symmetrical *relation* becomes "a relation that has no asymmetrical instances" (this negative characterization takes care of the *if-then* nature of the original definition, and is consistent with the fact that the empty relation is symmetrical). An asymmetrical relation becomes "a relation that has no symmetrical instances". And a non-symmetrical relation becomes "a relation that has at least one asymmetrical instance."

Defining "sometimes symmetrical and sometimes asymmetrical":

(8) The colloquial locution "On the domain H × H of all humans, sometimes *brother* is symmetrical and sometimes it's asymmetrical" can be defined as "The domain H × H can be partitioned into two non-empty subdomains such that *brother* is symmetrical on one and asymmetrical on the other."

If we let F be the set of female humans and M be the set of males, *brother* is symmetrical on M × M and asymmetrical on M × F ∪ F × M. (This isn't yet a partition of H × H, because we haven't included F × F. But since *brother* is the empty relation on F × F, it is both symmetrical and asymmetrical on that domain, so we could add F × F to either cell of the partition without changing the result.)

2.3. Alternative terminology

Not all texts use the terms *asymmetrical* and *non-symmetrical* in the way defined in (4).

(i) Gleitman *et al* use **asymmetrical** in the way that I defined *non-symmetrical.* Initially I thought their usage was a mistake, but I have discovered that both usages are common. Authors differ as to which term carries the meaning "not symmetrical".

I will continue to use *asymmetrical* as defined in (4), but it's important to be aware of the way Gleitman *et al* are using it, and the fact that that is also an accepted usage.

(ii) There are also two definitions of **non-symmetrical** in the literature.

(a) The definition in (4), "not symmetrical", can be called the "broad sense" of 'non-symmetrical'.

(b) The other, "narrow sense" defines non-symmetrical as "neither symmetrical nor asymmetrical".

There are advantages and disadvantages to both notions. The first choice is more intuitive and linguistically uniform: «non» means «not», and that holds also for *non-reflexive* and *non-transitive*. The second choice but not the first provides convenient terminology for a three-way partition of relations (on a specified domain) into symmetrical, asymmetrical, and neither.

3. What do "symmetrical", "non-symmetrical" apply to?

3.1. The "asymmetrical act of drowning someone".

In standard mathematical terminology, the properties "symmetrical" and "asymmetrical" apply to **binary relations** and nothing else. But in other parts of mathematics the properties may apply to geometrical figures, to certain algebraic structures, and to other mathematical objects. And in everyday English and in Gleitman *et al*'s article, those expressions have additional uses relating to acts, propositions, situations, as illustrated in (9).

(9) a. *asymmetric warfare*: warfare between two very unequal forces, with the two sides often using very different methods.

b. *an asymmetric power situation*: a situation involving two individuals or two states *a* and *b* in which *a* has much more power over *b* than *b* has over *a*.

c. *asymmetric interdependence*: a state of interdependence between two (or more) individuals or other entities, in which there are pairs *a*,*b* such that *a* is much more dependent on *b* than *b* is on *a*.

Gleitman *et al* begin their discussion of the relation between reciprocal structures with *each other* and reciprocal interpretations of intransitive symmetrical and non-symmetrical predicates, as in (2) and (3), with the observation that many obviously non-symmetrical predicates license the reciprocal construction. "Thus if John drowns Bill while Bill drowns John, we can say

(10) *John and Bill drown each other*." [2, p.326]

094-Partee_B:_Layout 1 18.05.2008 16:32 Page 609

They describe (10) as depicting a situation in which two individuals reciprocate the "decidedly asymmetric act of drowning someone." What is an "asymmetric act"?

3.2. In what sense is drowning an "asymmetric act"?

I can see three possible factors behind the idea that drowning is an "asymmetric act."

(i) In most cases where a drowns b, b does not drown a. If we make symmetry a graded notion with formal asymmetry and formal symmetry as the extremes, and rank the non-symmetrical relations on a scale according the proportion of aRb cases for which bRa also holds, then the relation aRb expressed by "a drowns b" is very close to the asymmetrical end of the scale. (Similarly, 'is a friend of' and 'is best friend of' may be non-symmetrical but close to the symmetrical end.)

(ii) We may be influenced by **geometrical symmetry**/asymmetry when we *picture* an act. Most imaginable ways in which someone drowns someone have the agent and victim in different positions – it's not a visually symmetrical picture. (But one can imagine a possible double murder where they manage to drown each other in a symmetrical-looking way.)

There is the related linguistic factor of Agent and Patient roles and how different they are. This is discussed by Gleitman *et al,* but in a different context, together with Figure-Ground changes, as involving changes in something other than actual semantic content. But semantics may be relevant: for transitive *drown*, the Agent and Patient roles are much more distinct than for, say *meet*, and there are no Agent or Patient roles at all with adjectival predicates like *similar* or noun predicates like *sibling* or *uncle* or prepositions like *near*.

(iii) The role of the **event argument** [5, 6] may be relevant. Active verbs have an event argument, which may not be present with adjectives or stative predicates. Suppose the basic argument structure of *drown* is as in (11).

(11) *drown*(*e,a,b*) : *e* is an event of *a* drowning *b*.

Then the question arises: can one and the same event be an event of a drowning b and b drowning a? This is a non-trivial question: compare debates about whether a buying event and a corresponding selling event are always/sometimes/never the same event. If the answer is "no", that is, if one can argue that they are never the same event even if they happen at the same time, that could be another basis for the intuition that drowning is an asymmetrical act.

3.3. What is the relation between 'drowned' and 'drowned each other'?

One of Gleitman *et al*'s insights is that a reciprocal sentence can in some sense turn an *asymmetrical* relation into a *non-symmetrical* one, i.e. can allow 'symmetrical instances' even if the core lexical predicate does not. How can we make formal sense of this intuitively appealing idea?

Consider sentence (10) again, here converted to the past tense for naturalness.

(10) *John and Bill drowned each other.*

And consider again Gleitman *et al*'s statement that "In (10), these two reprobates have reciprocated the decidedly asymmetric act of drowning someone." We have considered three possible bases for calling the act of drowning asymmetric. What might it mean then to say that John and Bill have "reciprocated" an asymmetric act?

First we note that the overt reciprocal construction certainly allows for two distinct events, as illustrated by the possibility of attributing different properties to the two events, as in (12).

(12) *John and Bill drowned each other by different methods.*

For a happier verb, and one for which it's easier to imagine two completely separate events, consider *rescue* in (13).

(13) *John and Bill rescued each other by different methods/ on different days.*

And as Gleitman *et al* note, even for a (relatively) symmetrical predicate like *kiss*, the reciprocal construction in (14a) allows for two distinct acts, unlike the intransitive (14b).

(14) a. *John and Mary kissed each other on different parts of their faces.*

b. *John and Mary kissed *on different parts of their faces.* [OK on an irrelevant iterative reading – many (mutual) kisses, in various places.]

094-Partee_B:_Layout 1 18.05.2008 16:32 Page 610

Returning to *drown*: Suppose that the meaning postulate in (15) is correct.

(15) $drown'(e,a,b) \rightarrow \neg drown'(e,b,a)$ for all a,b such that $a \neq b$.

This meaning postulate is a way of saying that the lexical *drown* relation is antisymmetrical relative to a fixed event e. (It's antisymmetrical, not asymmetrical, since a person can drown himself.)

Even if lexical *drown* is thus antisymmetrical, the *sentence* 'John drowned Bill' is consistent with the sentence 'Bill drowned John', because in a full sentence, the event argument is existentially quantified, as in (16).

(16) Possible: both $\exists e_1(\text{drown}'(e_1, a, b))$ and $\exists e_2(\text{drown}'(e_2, b, a))$

In other words, "John drowned Bill" and "Bill drowned John" can both be true by virtue of two different events, since each just says that *there is an event* of the given kind.

So if the *sentence* 'John drowned Bill' includes an existentially quantified event argument, that explains how the *sentence* can express a merely non-symmetrical relation, even if the core predicate *drown* inside it is asymmetrical or antisymmetrical. Thus we can formalize Gleitman *et al*'s idea that the reciprocal construction can turn an asymmetrical relation into a non-symmetrical one, one that can have "symmetrical instances."

4. Non-symmetrical predicates that pass the linguistic test for "symmetrical predicates"

Predicates of various syntactic classes that are classed as symmetrical predicates by Gleitman *et al* include the verbs *meet, kiss,* the adjective *similar,* and I will add the noun *sibling*.

The *linguistic* property considered as central by Gleitman *et al* for distinguishing the symmetric from the non-symmetric class is the test with *intransitives*, whether their meaning is or is not similar to the meaning of an overt reciprocal, as discussed in Section 1.

4.1. The observed meaning difference between reciprocal and plural intransitives with symmetric predicates.

Gleitman *et al* note carefully that even in the case of most of the predicates they class as symmetrical, such as *kiss*, the meanings of the intransitive-plural and the reciprocal sentences are not judged identical, only similar. This observed difference may also be explainable on the basis of the event argument plus the lexical semantics of the derived intransitive.

(17) a. *Susan and Bill kissed each other.*

b. *Susan and Bill kissed.*

(i) The adverbial test illustrated with (14a-b) supports the idea that there is just one event in the intransitive sentence, two in the overtly reciprocal sentence. But Gleitman *et al* raise the interesting question of how the structure and compositional semantic derivation of (17b) differ from that of a sentence with an intransitive variant of a non-symmetrical verb, such as *drown,* as in (3a), *John and Bill drown.*

(ii) First of all, let us grant that (17b) has a single event argument and describes the occurrence of a common event with two participants, and let us look at the structure of (17a). Appealing to analyses such as those in [7], [8], [9], we may argue that the overt reciprocal sentence (17a), while not derived from a conjunction of two sentences (we agree with Gleitman *et al* about that), nevertheless describes a non-atomic event, one that has parts that are subevents of the given event.

(iii) Returning to the intransitive (17b), I would argue that if it unambiguously posits a single event of mutual kissing, then the intransitive verb *kiss* needs to be considered a separate derived variant of the transitive *kiss*. Lexical rules for deriving several kinds of intransitive alternants of transitive verbs were proposed by Dowty [10]. Russian marks several sorts of derived intransitives (reciprocals, reflexives, unaccusatives) with the morpheme *sja.* English uses "zero-derivation" instead. The following is based on Dowty's rule, modified to include an event argument.

(18) transitive *kiss*: $\lambda y \lambda x \lambda e\ kiss'(e,x,y)$

inherently reciprocal $kiss_{\text{recip}}$: $\lambda X \lambda E\ [\forall y \leq_i X.\ \exists z \leq_i X.\ \exists e \leq E\ .\ kiss'(e,y,z)]$

This is to be read as follows: $kiss_{\text{recip}}$, the derived 'inherently reciprocal' version of *kiss*, is predicated of a single 'plural event' E and a single plural entity X, and $kiss_{\text{recip}}$ (E,X) says that for every atomic individual y who is part of X, there is a subevent e of E and an atomic individual z that is part of X such that e is an event of y kissing z.

It is likely that intransitive *kiss* has further lexicalized and has a more specific meaning than that, perhaps requiring that the ones who are kissing are not just kissing each other but kissing each other on the mouth. Dowty's lexical

094-Partee_B:_Layout 1 18.05.2008 16:32 Page 611

rules are proposed as semi-productive sense-deriving processes which can be followed by further lexicalization. The meaning suggested in (18) is a simple version of a reciprocal meaning; see the works mentioned above for more sophisticated suggestions for the semantics of reciprocals.

What is crucial is that it is possible to say that there is a single event with a single plural actant in the case of the intransitives, while characterizing that event in terms of subevents involving the corresponding transitive predicates.

4.2. Why does brother pattern as a symmetrical predicate?

We observed earlier that *brother* is non-symmetrical on the class of humans, and symmetrical only when restricted to the domain of males. Yet it patterns with symmetrical predicates in forming plurals as in (6). This seems to challenge one of the main conclusions of Gleitman *et al*, that "overwhelmingly often, symmetrical concepts are expressed by predicates marked with a special lexical feature. This lexical feature licenses a reading of noun phrase conjunction to express reciprocality of the relation between the nominal conjuncts. No asymmetrical [= non-symmetrical] concepts have this feature." (p. 365.)

A*symmetrical* is used in that paper as I use *non-symmetrical* (perhaps limited to non-symmetricals that are not "nearly symmetrical"). *Brother* is clearly non-symmetrical, since only about half the instances are 'symmetrical instances'. And yet it patterns like the fully symmetrical *sibling* and *cousin* and the nearly symmetrical *friend,* and unlike the nearly asymmetrical *uncle.*

Why does this happen with *brother* (and *sister*)? Is *brother* on the domain of males a separate concept? Is there a separate symmetrical lexical item *brother* distinct from the non-symmetrical one[1]? Or is it rather that with nouns, pluralization is able to pick out the symmetrical subrelation of the broader relation? I don't have an answer. One place I would look for ideas is in the work of Staroverov on conjoined relational nouns of the type *husband and wife, brother and sister* [11], since there is something intrinsically reciprocal about those predicates, and they are very close in meaning to *spouses, siblings* respectively.

5. Conclusions

A great deal of progress in semantics has emerged from studying areas where certain 'standard' logical notions do not seem to have a perfect fit with their nearest natural language equivalents, and finding better notions where they turn out to be needed. What I have argued in this paper is that there is an interesting field for research starting from apparent mismatches between logicians' definitions of symmetry and asymmetry and the way those notions are used in ordinary language, a field of research whose value is clear from the very fruitful insights in the pioneering work of Gleitman *et al*.

References

1. Tversky A. Features of similarity // Psychological Review, 1977. **84**. P. 327-352.
2. Gleitman L.R., Gleitman H., Miller C., Ostrin R. Similar, and similar concepts // Cognition, 1996. **58**. P. 321-376.
3. Partee B.H., ter Meulen A., Wall R. Mathematical Methods in Linguistics. // Dordrecht: Kluwer, 1990.
4. Birkhoff G., MacLane S. A Survey of Modern Algebra // New York: Macmillan, 1953.
5. Davidson D. The logical form of action sentences // The Logic of Decision and Action. Pittsburgh: Pittsburgh University Press, 1967. P. 81-95.
6. Kratzer A. Stage-level and individual-level predicates // The Generic Book. Chicago: The University of Chicago Press, 1995. p. 125-175.
7. Dalrymple M., Kanazawa M., Kim Y., Mchombo S., Peters S. Reciprocal expressions and the concept of reciprocity // Linguistics and Philosophy, 1998. **21** № 2. P. 159-210.
8. Eschenbach C. Semantics of Number // Journal of Semantics, 1993. **10**. P. 1-32.
9. Hackl M. The ingredients of essentially plural predicates // Proceedings of the 32nd North East Linguistic Society conference. Amherst, MA: GLSA, University of Massachusetts, 2002. P. 171-182.
10. Dowty D. Word meaning and Montague grammar. The semantics of verbs and times in Generative Semantics and in Montague's PTQ // Dordrecht: Reidel, 1979.
11. Staroverov P. Relational nouns and reciprocal plurality // Proceedings of SALT 17. Ithaca: CLC Publications, In press.

[1] Some of the participants in the 2007 colloquium at the University of Pennsylvania informed me that there are languages which have separate lexemes for 'brother [of a male]' and 'brother [of a female]', suggesting that the symmetrical subrelation is indeed a natural and salient concept.

095-Wilks:_Layout 1 18.05.2008 16:32 Page 612

ВИРТУАЛЬНЫЕ КОМПАНЬОНЫ ЧЕЛОВЕКА КАК НОВЫЙ ВИД ДИАЛОГОВОГО ИНТЕРФЕЙСА ДЛЯ БУДУЩЕГО ИНТЕРНЕТА

ARTIFICIAL COMPANIONS AS A NEW KIND OF DIALOGUE INTERFACE TO THE FUTURE INTERNET

Yorick Wilks *(yorickwilks@googlemail.com)*
University of Sheffield, UK.

В статье делается попытка связать будущее Интернета с новой, пока что относительно мало разработанной, технологией компьютерной реализации языка и речи. Концепцию, лежащую в основе этой технологии, я называю виртуальным компаньоном человека. Прежде чем обсуждать состав виртуального компаньона, необходимо упомянуть две технологии, не только потому, что они важны сами по себе, но также и потому, что относительно целей и достигнутых результатов каждой из этих технологий существует недопонимание. Конкретнее говоря, это

Языковые и речевые технологии

Агенты и семантическая сеть

К первой технологии имеет отношение представление Бернерса-Ли [Berners-Lee *et al*., 2001] о том, какие изменения предстоят Интернету. Именно для этого нового Интернета мы предназначаем виртуального компаньона – интерфейс человека и машины. Мы полагаем, что без такого компаньона пользоваться Интернет будет сложнее, а не проще. В конце статьи мы обратимся к семантической сети. Второе понятие – это агенты, которые из временных программных средств, способных, к примеру, обнаружить в Интернете дешевую веб-камеру, превратятся в постоянные элементы социального компаньона, способные взаимодействовать с пользователем в диалоговом режиме в течение долгого времени, усваивать потребности и предпочтения пользователя и в разговоре с ним сообщать большое количество жизненно важных данных.

Introduction

This is not a paper in social science, but rather in speculative technology: however, the underlying technologies exist already and I will briefly describe them, along with some account of the current debates over their consequences. The crucial move in the paper will be when, after describing Artificial Companions, real and possible, I go on to argue that they can be seen as links to the Internet, at least for vulnerable classes of people (the old, the young) but perhaps for all of us when faced with the coming torrent of information on the Internet, particularly information about ourselves.

Two component technologies

Before moving to describe the integration that constitutes the Companion, we must first mention two technologies, not only in their own right but because, in each case, there have been misunderstandings about their achievements and goals.

Language and speech technologies are, for our purposes, two closely related methods for interfacing to the Internet; the first by typing to it to ask a question or to ask it to do something, and the second by speaking and listening, for the same purposes. The two are related, in that speech technology normally decodes speech waves—i.e. what is said into a microphone – into some form like written text inside a computer, which is then analysed so as to be understood, with the effect that both spoken and written input end up being analysed in similar ways by what we are calling "language technology', which we can think of, loosely, as going from text to what it means.

The notion of a Companion

The paper introduces the notion of an Artificial Companion as a socially important paradigm for language and speech research in the next ten years: an intelligent and helpful cognitive agent which appears to know its owner and their

095-Wilks:_Layout 1 18.05.2008 16:32 Page 613

habits, chats to them and diverts them, assists them with simple tasks but makes no technical demands on them at all, and might be most suitable for vulnerable social groups like the young and the old. The paper also discusses current aspects of the overall speech and language research program that a Companion will need.

The technologies needed for a Companion are very near to a real trial model; some people think that Artificial Intelligence (AI) is a failed project after nearly fifty years, but that is it not true at all: it is simply everywhere. It is in the computers on 200-ton planes that land automatically in dark and fog and which we trust with our lives; it is in chess programs like IBM's Big Blue that have beaten the world's champion, and it is in the machine translation programs that offer to translate for you any page of an Italian or Japanese newspaper on the web.

And where AI certainly is present, is in the computer technologies of speech and language: in those machine translation programs and in the typewriters that type from your dictation, and in the programs on the phone that recognise where you want to buy a train ticket to, from among the four hundred or so British station names. But this is not a paper about computer technology any more than it is about robots, nor is it about philosophy.

Companions are not at all about fooling us as to their true natures, as in the Turing test scenario, because they will not pretend to be human at all: imagine the following scenario, which will become the principal one, running through this paper. An old person sits on a sofa, and beside them is a large furry handbag, which we shall call a Senior Companion; it is easy to carry about, but much of the day it just sits there and chats. Given the experience of Tamagochi, and the easily ascertained fact that old people with pets survive far better than those without, we will expect the Companion to be an essential lifespan and health improving object to own.

Other Companions are just as plausible as the Senior one, in particular the Junior Companion for children, that would probably take the form of a backpack, a small and hard to remove backpack that always knew where the child was. But the Senior Companion will remain our focus, not because of its obvious social relevance and benefit, possibly even at a low level of function that could be easily built with what is now available in laboratories, but because of the particular fit between what a Companion is and old people's needs.

Common sense tells us that no matter what we read by way of official encouragement, a large proportion of today's old people are effectively excluded from information technology, the web, the internet and advanced mobile phones because «they cannot learn to cope with the buttons». This can be because of their generation or because of losses of skill with age: there are talking books in abundance now but many, otherwise intelligent, old people cannot manipulate a tape recorder, which has too many small controls for them with unwanted functionalities. All this is obvious and well known and yet there is little thought as to how our growing body of old people can have access to at least some of the benefits of information technology without the ability to operate a PC or even a mobile phone.

After all, the needs of the elderly are real, not just to have someone to talk to, but to deal with correspondence from public bodies, such as councils and utility companies demanding payment, with the need to set up times by phone to be visited by nurses or relatives, how to be sure they have taken the pills, when keeping any kind of diary may have become difficult, as well as deciding what foods to order, even when a delivery service is available via the net but difficult in practice for them to make use of.

In all these situations, one can see how a Companion that could talk and understand on the phone, and also gain access to the web, as well as to process written text in email could become an essential mental prosthesis for an old person, one that any responsible society would have to support. But there are also aspects of this which go beyond getting information, such as having the newspapers blown up on the TV screen till the print was big enough to be read, and dealing with affairs requiring some degree of reasoning, like paying bills from a bank account.

We have talked of Companions as specialised computer agents for tasks as simple as using the web to find a supermarket's home delivery service for groceries. More interestingly, it may involve using the web to find out what has happened to their old school friends and workmates, something millions already use the web for. But we shall need some abstract notion of time lines and the coherence of life events on the web to sort friends and schoolmates from the thousands of other people with the same names.

The reasoning technologies we shall need to organise the life of a Companion's owner may turn out to be very same technologies needed to locate other individuals on the web and select them out from all the personal information about the world's population that fills up the WWW, given that the web is now not just for describing the famous but covers potentially everyone. Two of my friends and colleagues who are professors of computer science have some difficulty distinguishing, and maintaining a difference, between themselves on the web and, in one case, a famous pornography supplier in Dallas, and in another case a reasonably well known disc-jockey in Houston, all of whom are highly ranked by the Google algorithm [Page et al., 1998].

These problems – of sorting out who exactly web information is about – will soon become not just quirky but the norm for everyone, and what I shall want to argue later is that the kind of computer agency we shall need in a Companion, one that deals with the web for us if we are old or maybe just lazy, is in fact closely related to the kind of agency we shall need to deal with the web in any case as it becomes more complex. To put this very simply: the web will become unu-

095-Wilks:_Layout 1 18.05.2008 16:32 Page 614

sable for non-experts unless we have human-like agents to manage its complexity for us. The Internet/web itself must develop more human-like characteristics at its peripheries if it is to survive as a usable resource and technology: just locating a particular individual on the web, when a majority of the EU and US populations have a web presence, will become far more difficult and time consuming that it is now. If this argument is right, Companions will be needed by everyone, not simply the old, the young and the otherwise handicapped. It is going to be impossible to use of the web without its having some kind of a human face.

The notion of a Companion developed so far is anything but superhuman; it is vital to stress this because some of the public rhetoric about what companionable computers will be like has come from films such as 2001, whose computer HAL is superhuman in knowledge and reasoning. He is a very dangerous Companion, and prepared to be deceptive to get what he wants, which may be not at all what we want. Seymour Papert at MIT always argued that it was a total misconception that AI would ever try to model the superhuman, and that its mission was to model the normal, which was much the same as AI-pioneer John McCarthy's emphasis on the importance of common sense reasoning was on capturing the shorthand of reasoning, the tricks that people actually use to cope with everyday life. Only then would we understand the machines we have built and trained and avoid them becoming too clever or too dangerous. This same impetus was very much behind Asimov's Laws of Robotics, which set out high-level principles that no robot should ever break if it is to bring no harm to humans.

The difficulty with such principles is fairly obvious: if a machine were clever enough it would find a way of justifying (to itself) an unpleasant outcome for someone, perfectly consistently with acceptable overall principles. Doing that has been a distinctively human characteristic throughout history: one thinks of all those burned for the good of their own souls and all those sacrificed so that others might live. In the latter case, we are probably grateful for those lost in what were really medical experiments – such as the early heart transplants – even though they were never called that.

It will not be possible to ignore these questions when presenting Companions in more detail, and in particular the issue of where responsibility and blame may lie when a Companion acts as a person's agent and something goes wrong. At the moment, Anglo-American law has no real notion of any responsible entity except a human, if we exclude Acts of God in insurance policies. The only possible exception here is dogs, which occupy a special place in English law, at least, and seem to have certain rights and attributions of character separate from their owners. If one keeps a tiger, one is totally responsible for whatever damage it does, because it is ferae naturae, a wild beast. Dogs, however, seem to occupy a middle ground as responsible agents, and an owner may not be responsible unless the dog is known to be of "bad character". We shall return to this later and argue that we may have here a narrow window through which we may begin to introduce notions of responsible machine agency, different from that of the owners and manufacturers of machines.

It is easy to see the need for something like this: suppose a Companion told one's grandmother that it was warm outside and, when she went out into the freezing garden believing this, she caught a chill and became ill. One might well want to blame someone or something in these circumstances and would not be happy to be told that Companions could not accept blame and that, if one read the small print on the Companion's box, one would see that the company had declined all responsibility and had even got one to sign a document accepting this. All this may seem fanciful and even acceptable if one's grandmother recovered and the company gave the Companion a small tweak so it never happened again.

This story makes no sense at the moment, and indeed the Companion might point out with reason, when the maintenance doctor came round, that it had read the outside temperature electronically and could show that it was a moderate reading and the blame should fall on the building maintenance staff, if anywhere. These issues will return later but what is obvious already is that Companions must be prepared to show exactly why they said the things they said and offered the advice they did.

A Companion's memory of what it has said and done may be important, but will be used only rarely one hopes; though it may be necessary for it to repeat its advice at intervals with a recalcitrant user: "You still haven't taken your pills. Come on, take them now and I'll tell you a joke you haven't heard before". James Allen in Florida is already said to have modeled a talking companionable pill for the elderly!

The state of language and speech technology

How does this rather airy vision connect to the general state of R & D in speech recognition and natural language processing at the moment? My own belief is that most of the components needed for a minimally interesting Companion are already available; certainly the Companion is not particularly vulnerable to one major current technical weakness, namely the imperfect recognition rate of available Automatic Speech Recognition (ASR) systems. This, of course, is because a Companion is by definition dedicated to a user and so the issue of user-independent ASR does not initially arise, except when the Companion needs to make its own phone calls and understand what is said to it.

095-Wilks:_Layout 1 18.05.2008 16:32 Page 615

However, the Companion is not merely an application wholly neutral between current disputes about how best to advance speech and language systems, in part because it will surely need a great deal of representation of human knowledge and belief and therefore the Companion's development would seem to need overall approaches and software architectures that allow such representations and, ultimately, their derivation from data by machine learning. This last clause is very important because there has been a profound methodological shift in speech and language research in the last two decades. Before that, it was generally assumed that the knowledge of the world and of language that a machine intelligence required could be programmed in directly, the content being provided by the researcher's intuition. In the case of language, this assumption followed directly from Chomsky's [1972] approach to linguistics: that intuitions about the nature of language can be computed by rules written by experts who have intuitive knowledge of their (native) language.

All this has now turned out to be false: no effective systems have ever been built on such principles, nor (outside machine translation, perhaps) are they ever likely to be. The revolution that has replaced those doctrines holds that such knowledge, world or linguistic, must be gained from data by defensible (i.e. non-intuitionistic) procedures like machine learning.

In the late 1980's when symbolic natural language processing (NLP) was invaded by an empirical and statistical methodology driven by recent successes in speech processing. The shock troops of that invasion were the IBM team under Jelinek which developed a wholly novel statistical approach to machine translation (MT), one that was not ultimately successful [see Wilks 1994 for a discussion] but did better than anyone in conventional MT initially expected, and set in train a revolution in methodology in NLP as a whole.

Although the IBM team began without any attention to the symbolic content of linguistic MT, they were forced, by their inability to beat conventional MT systems in DARPA competitions, to take on board traditional linguistic notions such as lexicons, morphology and grammar, but they imported them not from intuitions but in forms such they could be learned in their turn and that fact was the ultimate triumph of their revolution.

The present situation in dialogue modeling – such as will be needed for a Companion – is in some ways a replay, at a lower level, of that titanic struggle. The introduction into ASR of so called "language models" – which are usually no more than corpus bi-gram statistics to aid recognition of words by their likely neighbours – -have caused some, like Young [2002] to suggest that simple extensions to current speech (ASR) methods could solve all the problems of language dialogue modeling.

Young describes a complete dialogue system seen as what he calls a Partially Observable Markov process, of which subcomponents can be observed in turn with intermediate variables and named (in order):

Speech understanding
Semantic decoding
Dialogue act detection
Dialogue management and control
Speech generation

Such titles are close to conventional for an NLP researcher, e.g. when he intends the third module as something that can also recognise what we may call the *function* of an utterance, such as that it is a command to do something and not a pleasantry. Such terms have been the basis of NLP dialogue pragmatics for some thirty years, and the interesting issue here is whether Young's Partially Observable Markov Decision Processes, are a good level at which to describe such phenomena, implying as they do that

The classic ASR machine learning methodology can capture the full functionality of a dialogue system, when its internal structures cannot be fully observed, even in the sense that the waves, the phones and written English words can be. The analogy with Jelinek's MT project holds only at its later, revised stage, when (as we noted earlier) it was proposed to take over the classic structures of NLP, but recapitulate them by statistical induction. This is, in a sense exactly Young's proposal for the classic linguistic structures associated with dialogue parsing and control with the additional assumption, not made earlier by Jelinek, that such modular structures can be learned even when there are no distinctive and observable input-output pairs for the module that would count as data by any classic definition, since they cannot be word strings but symbolic formalisms like those that classic dialogue managers manipulate.

The intellectual question of whether the methodology of speech research, tried, tested and successful as it is, can move in and take over the methodologies of language research may seem to many a completely arcane issue, like ancient trade union disputes in shipbuilding, say, as to who bored the holes and who held the drills. But, as with those earlier labour struggles, they seem quite important to the people involved in them and here, unlike shipbuilding, we have a clash of expertise but no external common-sense referee to come in and give a sensible decision.

095-Wilks:_Layout 1 18.05.2008 16:32 Page 616

Jelinek's original MT strategy was non/anti-linguistic with no intermediate representations hypothesized between speech input and speech output, whereas Young assumes roughly the same intermediate objects as linguists but in very simplified forms. So, for example, he suggests methods for learning to attach Dialogue Acts to utterances but by methods that make no reference to linguistic methods for this [known since Samuel et al., 1998] and, paradoxically, Young's equations do not make the Dialogue Acts depend on the words in the utterance, as all linguistic methods do. His overall aim is to obtain training data for all of them so the whole process becomes a single throughput Markov model, and Young concedes this model may only be for simple domains, such as, in his example, a pizza ordering system.

All parties in this dispute, if it is one, concede the key role of machine learning, and all are equally aware that structures and formalisms designed at one level can ultimately be represented in virtual machines of less power but more efficiency. In that sense, the primal [Chomsky, 1959] dispute between Chomsky and Skinner about the nature of the human language machine was quite pointless, since Chomsky's transformational grammars could be represented, in any concrete and finite case, such as a human being, as a finite state machine, of the sort espoused by Skinner.

All that being so, researchers nonetheless have firm predelictions as to the kinds of design within which they believe functions and capacities can best be represented, and, in the present case, it is hard to see how the natural clusterings of states that form a topic (such as, for example, how to build a jet plane, piece by piece) can be represented in finite state systems. It is equally difficult to see how the human ability to return in conversation to a previously suspended topic can be represented plausibly in such a way. But these are all matters that can be represented and processed naturally in well understood virtual machines above the level of finite state matrices [see Wilks et al. 2004].

There is no suggestion that a proper or adequate discussion of Young's views has been given here, only a plea that machine learning must be possible over more linguistically adequate structures than finite state matrices if we are to be able to represent, in a perspicuous manner, the sorts of belief, intention and control structures that complex dialogue modeling will need; it cannot be enough to always limit ourselves to the simplest applications on the grounds, as Young puts it, that « the typical system S will typically be intractably large and must be approximated ». In the end, the case put here may be no more than that the structures we use to represent our language, including to machines, must be comprehensible to us as humans.

The Semantic Web

Mention has been made earlier of the new form [Berners-Lee et al., 2001)] of the WWW as envisaged by Berners-Lee and colleagues to follow his original conception. This is a large topic and suitable for a separate paper [e.g. Wilks, 2006] and can be seen in two quite different ways: first, as the existing WWW but augmented by annotations on the items of all the texts it contains, so as to give more direct access to the meaning content of the texts.

On this view, the Semantic Web (SW) is an outgrowth of both language technologies, as described above and their notion of augmentations, which is partly inherited from initiatives in the Humanities (e.g. the Text Encoding Initiative, [see TEI]). These annotations could be seen as imposing a "point of view" on the SW, so that, for example, it might be possible to use the annotations to prevent me seeing any web pages incompatible with The Koran, and that might be an Internet-for-me that I could choose to have. But there is no reason why such an annotated web should necessarily, as some have argued [e.g. Nelson, 2005], impose a unique point of view. The technology of annotations is quite able to record two quite separate annotation data (as meta-data) for the same texts, and no uniformity of point of view is either necessary or desirable.

The second view of the SW, and one that Berners-Lee prefers, is that of an Internet whose content is accessible to Agents, partly through annotations and partly through data-bases whose semantics are well-known and understood. These agents [see e.g. Walton, 2006] operate on the Internet and provide services to customers, such as updating their diaries, finding cheap gas supplies etc. Such agents are therefore rather different from the concept of Companions, for they are transitory, and not designed for a permanent relationship with an owner based on extensive knowledge about the owner. One should note here, however, that contemporary work on the SW [e.g. Bontcheva *et al*., 2003, Ciravegna *et al*., 2003] has no need to choose between these two sources and functionalities I have distinguished, but rather seeks to combine both.

A third strand in the genesis of the SW is that of traditional AI itself and its long and honourable tradition of modeling reasoning, planning and knowledge representation. Some would argue the SW is no more than weaker form of AI which has sacrificed representational power to gain a system that works on a large scale.

Companions will draw on all these strands in the SW as well as that of the ECAs, or Embodied Conversational Agents [see e.g. Ruttkay and Pelachaud, 2004], although these have conventionally been conceived of not in language terms but of graphical, avatar, glance, expression and presence terms—i.e. with the emphasis on the visual, whereas the Companion is fundamentally an agent that establishes a relationship through talking, with all that entails in terms of politeness, emotion, personality and how those slippery but real concepts can be modeled in automata. But again, none

095-Wilks:_Layout 1 18.05.2008 16:32 Page 617

of these borderlines are firm: ECA: Companion, SW Agent: ECA, and most of the questions and technologies touched on in this paper apply not only to possibly permanent Companions but to a whole range of interactions with the Internet, from pseudo-boyfriends and – girlfriends, to recent results on determining and simulating author personalities in weblog texts [see Oberlander and Nowson, 2006].

In the coming decade the European Commission is planning huge investments in all these technologies under its Information Society Technologies (IST) program, and the edges of this research and the barriers to its advance should be much clearer during the coming Seventh Framework Programme [The COMPANIONS project will be supported 2006-2010 as the Integrated Project **IST-34434: Intelligent, Persistent, Personalised Multimodal Interfaces to the Internet].**

References

1. Ballim, A., and Wilks, Y. (1991) Artificial Believers: the ascription of belief. Lawrence Erlbaum, Hillsdale NJ.

2. Berners-Lee, T., Hendler, J., and Lasilla, O. (2001). The Semantic Web. Scientific American.

3. Bontcheva, K., and Cunningham, H. (2003) Information Extraction as a Semantic Web Technology: Requirements and Promises. Adaptive Text Extraction and Mining workshop.

4. Chomsky, N. (1959) Review of Skinner's Verbal Behaviour, Language 35: 26-58.

5. Chomsky, N. (1972) Language and Mind, Harcourt Brace, New York.

6. Ciravegna, F., (2003) Designing adaptive information extraction for the Semantic Web in Amilcare. In S. Handschuh and S. Staab, (eds.), Annotation for the Semantic Web, Frontiers in Artificial Intelligence and Applications. IOS Press.

7. Cole, R., Mariani, J., Uszkoreit, H., Varile, N., Zaenen, A., Zampolli, A., and V. Zue, (1998) Survey of the State-of-the-Art in Human Language Technology , Cambridge University Press.

8. Ferguson, C. H. (2005) What's Next for Google, In MIT Technology Review: http://www.technologyreview.com/articles/05/01/issue/ferguson0105.asp?trk=nl

9. FLIKR: http://www.flickr.com/

10. Memories for Life and Photocopains: http://www.memoriesforlife.org/

11. Oberlander, J. and S. Nowson, (2006) Whose thumb is it anyway?: Classifying author personality from weblog text.

12. http://www.hcrc.ed.ac.uk/~jon/papers/drafts/pc8.pdf

13. Page, L., Brin, S., Motawani, T., and T. Winograd (1998), The PageRank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project

14. Ramchurn, S. D., Huynh, D. and Jennings, N. R. (2004) Trust in multiagent systems. The Knowledge Engineering Review 19(1).

15. Ruttkay, Z., and C. Pelachaud, (2004) Evaluating Embodied Conversational Agents, Kluwer, Berlin.

16. Samuel, K., Carberry, S., and Vijay-Shankar, R. (1998). Dialogue Act Tagging with Transformation-Based Learning. In Proc. COLING98, Montreal.

17. TEI: http://www.tei-c.org/

18. Walton, C. (2006). Agents and the Semantic Web. Oxford, Oxford University Press.

19. Wilks, Y. (1994). Stone Soup and the French Room: the empiricist-rationalist debate about machine translation. Reprinted in Zampolli, Calzolari and Palmer (eds.) Current Issues in Computational Linguistics: in honor of Don Walker. Kluwer: Berlin

20. Wilks, Y., Webb, N., Setzer, A., Hepple, M., and Catizone, R. (2004) Machine Learning approaches to human dialogue modelling. In Kuppervelt, Smith (eds.) Current and New Directions in Discourse and Dialogue, Kluwer, Berlin.

21. Wilks, Y. (submitted 2006) The Semantic Web and the Apotheosis of annotation. Journal of Web Semantics.

22. Young, S. 2002. Talking to machines—statistically speaking, Proc. ICSOS02.

096-Narinjani:_Layout 1 18.05.2008 16:33 Page 618

ДИСКУССИОННАЯ ТРИБУНА

Этот раздел сборника предназначен для текстов дискуссионного характера. В отборе таких докладов Программный Комитет основывается прежде всего на их полемическом потенциале и актуальности для аудитории Диалога, а не научной или общественной оценке авторской позиции. Такие материалы могут не вполне соответствовать формату научного доклада на Диалоге, но зато поднимают темы, которые могут послужить основой дискуссий на тематически близких Круглых столах, устраиваемых во время конференции

БЕРМУДСКИЙ ТРЕУГОЛЬНИК: ВЗАИМОДЕЙСТВИЕ – КОММУНИКАЦИЯ – ОБЩЕНИЕ

BERMUDAN TRIANGLE: INTERACTION – COMMUNICATION – CONTACT

Нариньяни А.С. *(narin@aha.ru)*
ЗАО «ИнтеллиТек»

Обозначенный в заголовке треугольник представляет собой плохо определенную часть Системы знаний. С одной стороны, этой территории посвящено огромное число работ, с другой она остается слишком большой и недостаточно изученной для того, чтобы посмотреть на нее в целом и установить более конкретные границы между обозначающими ее понятиями. В значительной степени это связано с тем, что понятия эти – базовые и, следовательно, не слишком четко формулируемые. В частности, они тесно связаны с понятием информации, которое определялось неоднократно и по-разному, так что в данном контексте использовать его в качестве точки опоры не слишком продуктивно. В докладе делается попытка наметить хотя бы черновой набросок карты данного терминологического треугольника.

Здание нашего несколько искусственно созданного благополучия слишком легко может рухнуть, как только в один прекрасный день окажется, что при помощи нескольких магических слов, таких как информация, энтропия, избыточность..., нельзя решить всех нерешенных проблем.

К. Шеннон, статья «Бандвагон» [1]

Введение

Почему Бермудский треугольник? Потому, что эта аллюзия с моей точки зрения вполне уместна для обозначенной данной части территории Системы знаний. И границы ее не определены, и относящиеся к ней явления и процессы плохо пока поддаются комплексному исследованию. Однако пытаться делать это необходимо даже с риском заблудиться в густом тумане или утонуть в трясине многочисленных не слишком ясных понятий и вопросов.

Поэтому этот текст не претендует на то, чтобы хотя бы схематично наметить рамки области, которую можно было бы соотнести с темой *Модели Общения*. С самого начала приходится признать необозримую сложность проблемы и методологическую естественность замены общения коммуникацией, упрощающей объект рассмотрения на порядки, хотя и не делающей его при поиске подхода существенно более обозримым.

Тесная взаимосвязь понятий *коммуникация* и *информация* очевидна, однако, несмотря на обилие определений, продуктивно вписать их в наш треугольник не удается. Так что часть доклада тратится на обсуждение данной проблемы. Тем не менее, несмотря на остающуюся недоопределенность, делается попытка наметить набросок коммуникативного акта и процесса коммуникации, а также их связи между собой.

Объем доклада не позволяет ставить задачей полноту этого наброска, поскольку разным аспектам данной темы посвящены многие монографии, не говоря о сотнях статей.

Так что получилось то, что получилось.

096-Narinjani:_Layout 1 18.05.2008 16:33 Page 619

1. От общения к коммуникации

1.1. Необозримая сложность общения

1.1.1. Тема *общение* междисциплинарна и многомерна. Сегодня она представляется необозримой, поскольку расщепляется на бесчисленное множество проекций, включающих антропологию, этологию, социологию, психологию, семиотику, филологию, лингвистику, коммуникацию, технологию знаний и многие другие области исследований, а также их всевозможные пересечения. Не говоря о том, что каждая из этих проекций занята как теорией, так и приложениями, в которых *общение* так или иначе является либо основным, либо одним из основных объектом изучения.

Этот спектр так широк еще и потому, что *общение* является важным компонентом практически любой области человеческой деятельности, причем оно важно как в технологических процессах, так и в продуктах, ориентированных на «конечного пользователя». Эффективность и того и другого определяется эргономикой, а она – качеством учета лежащего в их основе *общения*.

Естественно, что необозримость этого пространства порождает сомнение в том, возможно ли выделить в нем то нетривиальное общее, которое с одной стороны заслуживает исследования per se, и с другой способно стать полезным для основной части проблем, так или иначе связанных с сутью общения.

1.1.2. Для ответа на подобный вопрос у науки есть традиционный метод:

(**i**) Собрать совокупность явлений, так или иначе относящихся к теме изучения.

(**ii**) Структурировать эту подборку, формируя такое понятийное пространство, в котором явления и их группы распределяются наглядно для подтверждения того, что это пространство достаточно отражает особенности объекта изучения и структуру его свойств, отличающих одни группы явлений от других;

(**iii**) Определить зависимости и взаимосвязи выделенных свойств, формирующих модель изучаемого объекта, которая (а) отличает его от явлений, похожих на исследуемые, но к ним не относящихся, и (б) позволяет рассматривать его обобщение в различных функциональных контекстах

Именно результат структуризации (ii) определяет достаточность сбора (i): ее успех доказывает, что собранная совокупность представительна для того, чтобы стать материалом для экспериментов по построению модели (iii).

Если же масштаб процедуры (i) - (iii) выходит за рамки обозримости, что мы и имеем в случае нашей темы, то естественным представляется шаг к упрощению изучаемого явления, существенно сокращающему размерность его пространства, но сохраняющему наиболее важные его составляющие. С тем, чтобы решения (ii) и (iii), если их удастся получить для упрощенного явления, могли бы послужить основой или каркасом тех же решений для исходного сложного.

1.1.3. Масштаб понятийного пространства *общение* побуждает искать в нем те компоненты, которые могут послужить точками опоры в поиске необходимого нам метода упрощения. Собственно поэтому нам и понадобилась система трех понятий, вынесенных в заголовок доклада.

Естественно, такая редукция сокращает пространство *общения,* условно говоря, «на порядок», но оставляет тот упомянутый выше каркас, который может обеспечить в дальнейшем восстановление – пусть частичное – этого пространства для разработки более полноценной модели. Тем более, что любые достаточно формализуемые компоненты общения могут рассматриваться как специальные виды информации, позволяя моделировать соответствующие классы общения как коммуникативное взаимодействие.

Для конкретизации обсуждения требуется уточнение смыслов, связываемых с понятиями *взаимодействие, коммуникация* и *общение* в данном докладе. Причем именно здесь приходится оговориться, что сопоставляемые им смыслы не претендуют на роль семантических определений для соответствующих русских существительных, - они вводятся ниже в рамках попытки неформального выделения подходящих компонентов Системы знаний, для которых эти слова используются всего лишь как *этикетки*, присвоенные данным компонентам.

1.1.4. Все сказанное подчеркивает чрезвычайную сложность обозначенного треугольника, что и побудило автора попытаться применить к нему метод редукции проблемы до некоторой обозримой сути.

Тем не менее, масштаб темы не позволяет пытаться втиснуть ее в узкие рамки доклада. Думаю, что если бы кому-то даже удалось составить структурированный перечень аспектов и составляющих предмета обсуждения, то подобный сухой реестр вряд ли мог представлять интерес сам по себе, поскольку каждый его пункт и общая структура требовали аргументированного обоснования.

Так как такой общей карты данного треугольника пока не существует и в ближайшей перспективе не предвидится, мне придется ограничиться рассмотрением небольшого числа почти наугад выбранных тем. Некоторые – может быть и большая часть - не будут оригинальными, однако исключение их не позволило бы добиться хоть какой-то наглядности картины. Этот эксперимент проводится мной в надежде на то, что подборка, каждый компонент которой достаточно известен, может в сумме дать что-то новое.

1.2. Реперные точки пространства общения

096-Narinjani:_Layout 1 18.05.2008 16:33 Page 620

1.2.1. Определим соответствующие термины, относящиеся к ключевым для нашей темы компонентам Системы знаний (СЗн).

Воздействие: базовый акт СЗн, связывающий две сущности (два агента) – субъект и объект, и, соответственно, инициацию воздействия и реакцию на него; любые сущности СЗн так или иначе определяются и реализуются спектром осуществляемых ими воздействий.

Взаимодействие: отношение, предполагающее участие как минимум двух воздействующих друг на друга реальных и\или виртуальных агентов, и допускающее, но не требующее наличия информационной составляющей, поскольку оно может относиться к любой проекции СЗн, - механике, химии, квантовой физике, функционированию автономных клеток, простейших и сложных организмов, любых технических устройств и т.д.

Коммуникативный акт (КА): акт воздействия, необходимым компонентом которого является информационная составляющая; включает акты порождения и восприятия информации.

Коммуникация: взаимодействие, агенты которого участвуют в обмене коммуникативными актами.

Акт общения: коммуникативный акт, включающий эмоциональную составляющую.

Общение: коммуникация, включающая акты общения.

Эти достаточно упрощенные определения будут ниже использоваться для детализации обсуждаемого пространства и, в частности, их самих.

1.2.2. Из этих определений ясно, что среди трех перечисленных в заголовке понятий *взаимодействие* принимается как наиболее общее. По отношению к нему *коммуникация* – определяется как более частное, так как представляет собой особый вид взаимодействия, необходимым элементом которого является выделенная информационная составляющая.

И, наконец, *общение* – это тоже коммуникация, однако в общем случае не сводящаяся к обмену информацией и не ограниченная только ею. Поскольку общение может быть чисто эмоциональным и\или физиологическим, то будем считать, что в нем информационная составляющая хотя и всегда присутствует, но иногда как латентный «технологический» компонент, а не как ключевая составляющая, определяемая функциями высшей нервной деятельности.

Таким образом, кроме всех прочих подходов к изучению общения, естественным в рамках методологии редукции (п.1.1) представляется упростить рассмотрение *общения* переходом к *коммуникации*, которую интерпретировать *как технический обмен информацией*, допускающий исключение большей части составляющих, делающих общение предметом изучения перечисленных в первом абзаце п.1 наук.

Поскольку пространство коммуникации также многомерно, нам придется пытаться обозначить некоторое множество измерений, что в рамках доклада возможно только самым грубым их структурированием, в основном, разделением на две или несколько категорий.

1.2.3. Далее мы будем различать коммуникацию:

– *Прямую,* в ее обычном смысле, как систему информационных обменов, реализуемых через коммуникативные акты взаимодействующих между собою агентов; это предполагает коммуникацию в реальном времени или, по крайней мере, с использованием традиционных носителей информации (текст, аудио- и видеозапись, и т.п.) и

– *Обобщенную*, так же включающую обмен коммуникативными актами между участвующими в ней агентами, но, тем не менее, выходящую за рамки прямой, поскольку все ее составляющие относятся к коммуникации в максимально расширенном, обобщенном смысле.

Основное содержание статьи относится к прямой коммуникации, хотя местами в рассмотрение приходится включать и обобщенную для сохранения пропорций намечаемой системы понятий.

В дополнение к этой дихотомии далее мы будем должны рассматривать и другие проекции пространства коммуникации, разделяемые на соответствующие классы и категории, помогающие нам как-то ориентироваться в этом многомерном пространстве.

Соответственно любая классификация в докладе остается неформальной, поскольку относится к тому уровню понятий, на котором формализация либо невозможна, либо затруднена. Для нашей задачи попытки в этом направлении были бы вряд ли продуктивны, тем более что эти понятия будут до известной степени уточняться по ходу обсуждения.

1.3. Границы треугольника

1.3.1. Соответственно, коммуникативные системы условно можно разделить на:

– *Детерминированные*, в которых порядок взаимодействия агентов и функции КА определены точно и могут быть описаны формально (например, технические системы, жесткие регламенты и ритуалы и т.п.).

096-Narinjani:_Layout 1 18.05.2008 16:33 Page 621

– *Естественные* (живые), ход которых определяется не только правилами и процедурами, но и плохо формализуемыми составляющими, включающими физиологические ощущения и чувства, подсознательные и осознаваемые эмоции, рефлексию, инициативу и свободу воли их агентов.

Если общение – это коммуникация плюс включение во взаимодействие перечисленных выше составляющих живых систем, то именно их влияние является необходимым контекстом общения.

Понятие *информация* и ее связь с плохо формализуемыми составляющими является здесь одним из центральных, поэтому ниже ей будет уделяться значительное внимание.

1.3.2. Неформальность нашего подхода позволяет ограничиться при разделении используемых в ней понятий «волюнтаристскими» решениями. В частности, можно опереться на различие агентов взаимодействия по *форме реакции* на него, которая с возрастанием сложности разделима на следующие классы:

– *Базовая,* управляемая фундаментальными законами природы (физики, химии и т.п.),

– *Рецепторная,* переводящая внешнее воздействие на уровень, определяемый более сложными процессами, не сводящимися к фундаментальным, но в изученных случаях достаточно четко определяемых хотя бы в формате «черного ящика»,

– *Физиологическая,* включающая в восприятие нервную систему,

– *Эмфатическая,* включающая эмоциональную составляющую,

– *Интеллектуальная,* к которой отнесем все уровни, включающие такие составляющие как анализ, прогноз, учет опыта, влияние психологических, социальных и других факторов, и т.п.

1.3.3. Приведенная классификация остается неформальной, но дает возможность, хотя и достаточно условно, уточнить границы нашей схемы. В частности, оно позволяет:

– Отделить в системах взаимодействия агенты с базовой реакцией (биллиардный шар) от тех, у которых взаимодействие включает аппарат технических датчиков \ сенсоров (устройство управления) или биологических рецепторов (инфузория). Таким образом, базовая реакция отнесена к компонентам взаимодействия без участия информации.

– Отнести два последних типа к уровню, связанному с формированием общения.

– Что касается второго и третьего уровней, то к коммуникации их можно отнести без сомнений только в том случае, когда уровень, на который переводится ими внешнее воздействие, и функции аппарата обработки поддаются хотя бы какой-то формализации.

Таким образом, простейшее взаимодействие сводится к процессу, участники которого обмениваются *базовыми* актами воздействия, не включающими передачу информации. В этом случае акт воздействия не требует участия в процессе аппарата восприятия и мы исключаем этот простейший акт как упрощение понятия коммуникативного акта или, тем более, акта общения. Разве что на уровне метафоры.

1.4. Границы информации

1.4.1. Очевидно, что сомнения К.Шенона, отраженные в эпиграфе, побуждаю к попытке уточнения границ понятия *информация* в контексте нашего треугольника.

В дополнение к проведенному выше делению коммуникативных систем на детерминированные и естественные, можно разделить их на *неорганические* и *органические,* исключая из тех и других базовые. Это позволяет детализировать нашу классификацию.

В частности, к неорганическим коммуникативным системам относятся технические информационные системы, т.е. искусственные объекты, созданные для обработки различных форм информации.

В данном случае даже данное определение далеко от точности, поскольку в широком спектре аналоговых устройств – прежде всего, наиболее простых – трудно четко отделить информационные процессы от базовых. Те же весы, оставаясь простым механизмом, выполняют функции перевода базового воздействия в числовую информацию.

Что касается органических систем, то эта категорию можно условно разделить на следующие классы:

A. Простые и сложные органические молекулы, объекты органической химии.

B. Простейшие одноклеточные и многоклеточные организмы,

C. Организмы с вегетативной нервной системой,

D. Организмы с высшей нервной деятельностью.

Если класс A исключается из коммуникативных систем, как относящийся к базовым, то C и D – заведомо оперируют информацией высокого уровня. Что касается даже самых простых организмов класса B, то и в них функционируют комплексы взаимосвязанных сложных структурных органических молекул, взаимодействующих друг с другом и средой. Взаимодействие этих комплексов внутри организма вряд ли можно свести только к базовому, откуда следует, что уже на этом уровне явно имеет место обмен какой-то специальной, возможно, очень простой, формой информации. Но еще на шаг выше, например, в функционировании генных

096-Narinjani:_Layout 1 18.05.2008 16:33 Page 622

механизмов, прорисовывается сложная многомерная информационная система, реализованная природой на уровне биохимических процессов.

Еще более очевидно использование информации в процессах деления клеток и их коллективного взаимодействия со средой.

1.4.2. Сказанное выше побуждает нас рассмотреть ту же проблему с разделением взаимодействия на две категории *внутреннее* и *внешнее,* где *внутреннее* является элементом процесса взаимодействия между компонентами (подсистемами) объекта.

Поскольку любой реальный агент состоит из компонентов, а те, в свою очередь, также представляют собой системы, и так вплоть до атомного уровня и далее, то компоненты нижнего уровня взаимодействуют на основе базовых законов квантовой механики.

Этажом выше на молекулярном уровне взаимодействие идет в рамках законов физической химии, и с дальнейшим повышением сложности молекулярных конструкций и подсистем механизмы взаимодействия усложняются настолько, что относить их к базовым становится все трудней.

Если агент представляет собой сложный объект, участвующий в коммуникации или тем более в общении, то где-то между уровнем внешнего многообразного интеллектуального взаимодействия и самыми нижними этажами базового мы сталкиваемся с terra incognita, где на текущем уровне рассмотрения невозможно определить, к какому классу относятся механизмы внутренних подсистем.

Даже на нижних уровнях взаимодействия граница между ними обозначается очень условно. Например, между этажами квантовой физики и химии граница определяется тем, что количественная и качественная необозримость физического уровня заставляет переходить на другой – химический - уровень моделирования.

1.4.3. Тот же компьютер, который для нас безусловно относится к классическим системам обработки информации, вместе с тем «внутри» весь находится на одном уровне электромагнитных процессов. Однако, поскольку на этом нижнем этаже масштаб системы необозрим, мы вынуждены подниматься с уровня на уровень: элементарные логические устройства, функциональные модули, технологические компоненты, основные блоки, переходя от электроники к логике и информатике.

Пожалуй, самая наглядная иллюстрация, – это сам человек, представляющий собой неисчислимое множество атомов (порядка 10^{27}), которое, несмотря на его абсолютную необозримость, сохраняет физическую и функциональную целостность на всех многочисленных десятках своих системных этажей.

В заключение этой неразрешимой шарады приведем в качестве простого, но наглядного примера обычные ходики. Базовый уровень взаимодействия их компонентов, как очевидно, примитивно механический. Что же касается функции часов в процессе коммуникации, то они явно постоянно генерируют *пассивный фоновый* КА (см. п.2.2), сообщая другим агентам текущую оценку времени.

Таким образом, наши попытки установить границу между переработкой информации с одной стороны и чисто физическими, химическими и прочими актами воздействия с другой, каждый раз подтверждают, что она очень условна.

Похоже, что здесь работает известный принцип тесной взаимосвязанности объекта наблюдения и наблюдающего. Хотя операции восприятия и переработки информации базируются на тех же природных базовых процессах, но когда их изучения поднимаются на уровни, где процедуры взаимодействия становятся слишком сложными для описания их как суперпозиции базовых, они становятся для нас очевидными или латентными коммуникативными актами.

1.4.4. При этом многие вопросы остаются для нас открытыми, в частности:

– Можем ли мы все то, что сложнее базовых процессов, связывать с обработкой информации или там имеет место нечто, выходящее за эти рамки? Например, является ли эта обработка необходимым, а, главное, ключевым, компонентом физиологических ощущений? Если они – или хотя бы одно из них – есть у той же инфузории, то можно ли считать это формой восприятия и переработки информации? И если да, то может ли данный факт считаться достаточным для возможности рассматривать инфузорию потенциальным участником простейшей формы коммуникации, способным воспринимать как информацию базовые внешние факторы, - температуру, давление, химический состав среды и т.п.?

– Как ответ на предыдущий вопрос влияет на постановку того же вопроса по отношению к вирусам, эритроцитам и другим простейшим «живым» агентам, о физиологических «ощущениях» которых говорить трудно, но реакция которых на внешние воздействия и организация внутренних процессов слишком сложны, чтобы отнести их к чисто техническим?

Выше было обозначено, что воздействие может быть *базовым, сенсорным, нервным, психологическим, эмфатическим* и\или *интеллектуальным*. Однако в этом ряду границы между базовым взаимодействием, коммуникацией и общением далеко не очевидны.

С одной стороны, с взаимодействием любой сложности биллиардных шаров можно считать все ясным. С другой, восприятие информации системой, претендующей по сложности на интеллектуальность, далеко не гарантирует необходимость остальных составляющих: она может получать информацию в «готовом» виде, не

096-Narinjani:_Layout 1 18.05.2008 16:33 Page 623

требующем даже сенсорных рецепторов. При этом и с искусственной интеллектуальной системой далеко не все так просто, если в нее включены модели психологической и\или чувственной составляющей, сложность которых по спектру реакций может превосходить в какой-то области палитру реакций жука, мыши, а то и «простого» человека.

Таким образом, очевидно, что для нашей задачи замещение общения коммуникацией последняя может также оказаться избыточно сложной, для уточнения чего нам приходится вернуться к более детальному определению *коммуникативного акта*.

2. Коммуникативный акт

2.1. КА от микро до макро

2.1.1. В качестве коммуникативного акта может рассматриваться передача информации любой формы и масштаба от простейшего жеста до книги, телесериала или даже до таких масштабных реалий как та или иная цивилизация в целом. В конце концов, любое явление каждой эпохи, в том числе и сама эта эпоха, является коммуникативным актом, который будет как-то воспринят следующими поколениями. Хотя он обычно им не адресуется, и доходит до них либо далеко не полностью, либо «теряется по дороге».

Стоит заметить, что каждая культурная эпоха – это внутренний диалог, не осознающий того, что является наряду с этим и КА, отправляемым в будущее. Конечно, есть исключения, но их не так уж много, - например: пирамиды и большие храмы всех времен, строившиеся на века, мировая революция, тысячелетний рейх, рукописи в стол и даже письма комсомольцев времен застоя, закладывавших в капсулы послания к молодежи следующих поколений [2].

2.1.2. Для расширения нашей темы стоит отметить, что КА может быть *активным* и\или *пассивным*.

Первая категория включает все КА, генерируемые для восприятия *здесь* и *сейчас*, т.е. в конкретных, так или иначе ограниченных текущим контекстом, рамках информационного процесса. Сюда же относятся и кратковременные внешние процессы, воспринимаемые как поступающая через органы чувств информация о текущих происходящих вокруг событиях.

К категории пассивных отнесем те КА, которые не ориентированы на восприятие on line. Например, пассивными КА можно считать как артефакты с длительным временем существования (вывеска, витрина, здание, город, т.д.), так и элементы реальности, воспринимаемые как компоненты внешней ситуационной информации (дерево, закат, зима и т.п.).

Понятно, что значительная часть коммуникативных актов являются активными и пассивными одновременно: книга, написанная в своем временном контексте, может стать КА на века и тысячелетия (классика или раритет)

2.1.3. С этой точки зрения любой доступный для наблюдения базовый акт воздействия есть КА, поскольку система восприятия получает информацию о факте этого акта, его агентах и следствиях. Таким образом вся окружающая обстановка представляет собой хаос или симфонию (все зависит от системы восприятия) необозримого числа КА, формирующих для нас панораму пассивных и активных компонентов текущей картины реальности, воспринимаемой через Систему знаний.

Для того, чтобы элемент окружающей обстановки стал компонентом коммуникации, отраженный им свет (или излучаемый, если он является источником) должен быть *воспринят как информация*, иначе это активное или пассивное потенциальное сообщение останется базовым актом. Важно при этом, что смысл КА определяется не только его источником, но и системой восприятия, принимающего его как информацию.

Фоновые элементы участвуют в коммуникации, посылая системам восприятия действующих агентов информацию о самом факте своего существования и о событиях имеющих место через их взаимодействие. Они могут никак не использоваться в процессе коммуникации активных агентов и даже оставаться за рамками их восприятия, но могут играть и ведущую роль. Например, когда некто, мирно говорящий с соседями о погоде, неожиданно видит горящий бикфордов шнур, ясно, что тема, стратегические задачи и форма текущей коммуникации с этого момента радикально меняются.

Остается отметить разницу между упоминавшимися выше часами и горящим бикфордовым шнуром: первые созданы для формирования и передачи пусть специфической (время), но информации, в то время как второе обычно для передачи информации не предназначено. Хотя в конкретном случае все может быть наоборот: часы могут быть часовым механизмом взрывного устройства (или не работать, пассивно сообщая о своем бездействии), а горящий шнур окажется демонстративным актом угрозы, требующим мгновенной реакции.

2.2. Классификация КА

Оставляя мега- и макро- коммуникативные акты в стороне и ограничиваясь типами КА, относящимися к текущей деятельности, мы можем очень условно разделить их на *фоновые, разовые, локальные* и *структурные*.

096-Narinjani:_Layout 1 18.05.2008 16:33 Page 624

– *Фоновый* КА: акцентированный активный или пассивный элемент контекста, передающий информацию об особых функциях агента или обстановки: деталь или особенность одежды, предмет в руках, стиль или деталь здания, вывеска и т.п., все, что может играть свою – иногда важную – роль в процессе коммуникации.

– *Отдельный* КА: коммуникативное действие между моментом начала процесса передачи информации до его завершения.

– *Сложный* КА: связанные между собой отдельные КА, представляющие комплексный обмен информацией между двумя или более коммуникантами, ориентированный на решение одной или нескольких взаимосвязанных задач. Границы сложного КА определяются вложенными в него отдельными КА, - открывающим этот обмен и завершающим его. Таким образом, сложный КА – это содержательно выделяемый «сеанс» коммуникации.

– *Структурный* КА: комплекс связанных между собой сложных КА, представляющий собой обмен информацией между агентами в рамках одной коммуникативной системы, решающей одну или несколько взаимосвязанных задач. Границы структурного КА ограничены сложными КА, открывающим этот обмен и завершающим его.

КА любого из перечисленных классов может быть прерванным и\или незаконченным, причем само внешнее или авторское прерывание также является коммуникативным компонентом КА, своего рода микро-КА.

Данная классификация условна, поскольку трудно, а во многих случаях и невозможно, определить границы даже отдельного КА, не говоря о более сложных, которые могут иметь многоуровневую структуру.

2.3. Обязательные составляющие КА

2.3.1. Коммуникативный акт мы определили как акт воздействия, включающий информационную составляющую. Ограничимся активным КА, в котором участвуют:

– *Субъект* (автор, источник), осуществляющий акт генерации (формирования) информации, ориентированный на конкретный или обобщенный объект воздействия, - адресата или «виртуального» получателя КА, например, *всем – всем – всем*.

– *Агент* восприятия информации, конкретный, виртуальный или обобщенный получатель, способный сознательно или подсознательно, адекватно или неадекватно воспринять это воздействие как информацию.

При прямой коммуникации источник КА является его автором, а получатель в общем случае совпадает с адресатом. Хотя от этой стандартной схемы возможны различные девиации, например:

- получатель принимает за КА некоммуникативный акт воздействия (принимает шум за речь), или
- автор адресует КА неадекватному объекту («*Дорогой многоуважаемый шкаф*), и т.п.

Понятно, что для того, чтобы стать полноценным участником коммуникации, надо принять информацию, переработать ее и отреагировать на нее, в частности в форме информации, ориентированной на автора входного КА или другого \ других агентов.

2.3.2. При этом в самом КА можно выделить пять его обязательных составляющих:

– Материально-энергетическая составляющая, являющаяся носителем воздействия, доступного органам чувств живого организма или приборам восприятия технического устройства; такими носителями могут быть визуальная и\или акустическая среда, технические электромагнитные частоты, изменения параметров излучения, биохимические процесс в организме, и т.п.

– Канал, конкретизирующий форму передачи и\или приема сообщения; например, конкретный телевизионный или радио канал, непосредственное восприятие звука или прямое видение источника визуальной информации, и т.п.

– Форма «кодировки» информации КА в рамках возможностей канала: речь, видео и\или аудио формат, файл doc, азбука Морзе, выстрел из пушки, и т.п.

– Текст КА, т.е. представление информации на языке коммуникации,

– Содержание сообщения КА.

Конечно, это - очень упрощенная схема, в которой каждый пункт представляет собой пакет более частных составляющих. Кроме того, в одних случаях некоторые составляющие могут совпадать, - например, текст и содержание (прямым текстом передается пароль), в других - содержание может быть достаточно сложным и «многослойным», в третьих – элементы формы текста могут быть частью его содержания и т.д.

2.4. Компоненты обмена сообщениями

Для реализации законченного КА как элемента коммуникации необходим не только акт посылки информации, но и ее приема. Другими словами, полноценный участник коммуникации должен обладать полным набором функциональных компонентов обмена сообщениями, включая аппараты:

i. Восприятия, переводящий процесс чисто материально-энергетической составляющей на другой - информационный - уровень. Для нас в данном случае неважно, является ли аппарат восприятия компонентом

096-Narinjani:_Layout 1 18.05.2008 16:33 Page 625

биологического или технического объекта. Существенно лишь то, что этот уровень определяет новое качество, которое превращает данный акт воздействия в коммуникативный.

ii. Переработки полученной информации, извлечения ее содержания,

iii. Формирования реакции на КА, которая может определяться сочетанием широкого спектра составляющих: от безусловного или условного рефлекса до сложной интеллектуальной деятельности, вырабатывающей ответ на внешнее воздействие. Важно, что этот ответ может относиться к коммуникативному отзыву на КА и\или быть элементом реакции на воздействие внешней среды.

iv. Организация формы ответного КА*, т.е. превращения его информационного содержания в коммуникативный акт, определяемый выбором:

– формы материально-энергетического посыла,

– канала, конкретизирующего форму передачи,

– языка коммуникации, т.е. «кодирования» информации КА* в рамках возможностей канала,

– «текстом» КА*, т.е. представлением содержания реакции на языке информационного обмена.

Первые три компонента выбора могут быть «встроены» в аппарат передачи коммуникативного акта, но формирование последнего определяется его содержанием.

Наличие аппаратов i и ii обеспечивает функции приема информации, а аппаратов iii и iv – ее передачу.

Если какие-то из перечисленных компонентов обмена сообщениями у участников отсутствуют, то коммуникация превращается в разнородное взаимодействие, в которое могут быть включены элементы генерации или приема информации, и даже обмена ею, но связанного полноценного процесса коммуникации не складывается.

3. КА в контексте коммуникации

3.1. Импульс к формированию КА

3.1.1. Создание коммуникативного акта начинается с внутреннего импульса (побуждение, толчок, стимул) к его формированию, требующего определения *с какой целью, кому, что и как сообщить.* Очевидно, что перечисленные четыре составляющие частично или полностью зависят от:

– функций (специализации) коммуниканта,

– его представления о текущем ходе процесса взаимодействия в целом или той его части, в которой он принимает участие,

– уровнем его «знаний» о семантике и прагматике процесса и т.п.

При этом перечисленные составляющие могут:

– либо контролироваться системой коммуникации (ее центром и\или регламентом и\или «анамнезом» процесса и т.п.)

– либо формироваться автономно агентом коммуникации,

– либо определяться ими совместно с тем или иным разделением функций принятия решений, формирующих данный КА.

При этом импульс может быть *случайным \ регулярным \ «программируемым»* процессом внутри агента и\или в коммуникативной системе.

3.1.2. Сформулировать информацию импульса «буквально» возможно только в полностью регламентированной системе, где указанные составляющие импульса определяются достаточно точно. В общем же случае речевого общения импульс, как правило, не вербален по сути, поскольку на его «озвучивание» влияет масса факторов, достаточно сложно взаимосвязанных.

Таким образом, в сложных системах коммуникации смысл созданного КА может оказаться не совсем темт, к которому побуждал начальный импульс и который собирался передать автор КА. Каждая из составляющих импульса могла иметь несколько взаимосвязанных альтернатив, с одной стороны упорядоченных по приоритету, а с другой ограниченных текущими возможностями агента. И это сочетание решений может быть далеко не оптимальным по отношению к требуемым и\или желательным.

Оценить то, что получилось по отношению к исходному побуждению, может только сам автор КА, поскольку:

(а) только автору - и то далеко не полностью - известен (понятен) тот импульс, который инициировал КА (если этот импульс в законченной форме не пришел извне), и

(б) автор является одновременно и воспринимающим данного акта, поскольку в случае интеллектуальной коммуникации должен учитывать конкретику сформированного им КА. А в случае речевого общения КА часто по сути обращен автором к самому себе, а совсем не к адресату (или к адресату в меньшей степени).

[1] Мы рассматриваем тут именно эту фразу, а не максиму *мысль изреченная есть ложь*, имеющую несколько смыслов и плохо подходящую для данного примера.

096-Narinjani:_Layout 1 18.05.2008 16:33 Page 626

3.1.3. В докладе [2] был раздел, специально посвященный иллюстрации высказывания *слово изреченное есть ложь*[1] Причем в данном случае речь не шла о намеренном обмане, а о пошаговом разборе процесса генерации и восприятия речевого акта (РА), результат которого в общем случае может быть достаточно далек от намерений автора.

Существенно упрощенная схема передачи речевого акта состояла в докладе из четырех фаз:

- то, что говорящий хотел отразить в РА,
- то, что ему удалось вербализовать,
- то, что адресат «получил на вход»
- то, что адресат извлек из РА.

При этом та часть содержания, которая была воспринята адресатом из всего информационного импульса, послужившего причиной формирования РА, может оказаться искаженной и незначительной. Причем легко представить ситуации полного непонимания или восприятия РА в смысле, противоположном исходному. Особенно это ясно при общении носителей разных языков, разных социальных слоев и, тем более, разных культур.

Кроме всего прочего, этот пример напоминает и о еще одной ключевой проблеме коммуникации: поскольку КА в любой форме связан с передачей информации, то для восприятия ее содержания необходимы какие-то оценки ее истинности. Не абсолютной, о которой говорить в данном случае не имеет никакого смысла, а об отношении содержания полученного КА к целям и задачам того или иного уровня, того или иного участника, в том или ином коммуникативном контексте, в отношении тех или иных критериев, и т.п.

Очевидно, что все затронутые выше проблемы являются объектами широких исследований в разных областях, так что эту тему мы оставляем вне доклада в надежде вернуться к ней в будущем.

3.2. Кому и как

3.2.1. Естественно, что в решение автора, *как сообщить,* входит определение канала передачи, который может быть жестко задан или выбираться из некоторого множества возможностей.

Выше соответствующее пространство выбора уже рассматривалось, как сочетание среды передачи информации, канала, конкретизирующего форму передачи сообщения и языка коммуникации в рамках возможностей канала.

В «обычной» коммуникативной системе такого выбора не требуется, поскольку эти параметры фиксированы. Но в общем случае, - например, в условиях, рассчитанных на устойчивость к тем или иным помехам (техническим, враждебным и т.п.), система часто включает возможность изменения среды, канала и\или языка коммуникации. Причем это может происходить автоматически или требовать выбора настройки на основе взаимодействия с адресатом и\или системой. Например: *слышно тебя очень плохо, пошли СМС.*

Соответственно, в варьировании языка и средств защиты часть возможностей выработала природа: мимикрия, запутывание следов, различные метки и т.п. Но это множество на порядки уступает сложности и разнообразию форм связи, созданных человеком: тысячи естественных языков и диалектов, приемы шифровки, эзопово иносказание, оттенки интонации, мимики и жестов, и т.п.

В общем случае, смена средств языка или их сочетание связаны не только с проблемами защиты информации, но и с расширением палитры передачи смысла. Так даже в печатном тексте применяется игра шрифтами и форматами для более наглядной передачи различных планов содержания. Тем более это относится к просодии, мимике и жестам при устной речи.

3.2.2. Адресат КА может быть *конкретным, всеобщим* или *неопределенным.*

– К первой категории относится реальный участник (группа участников) процесса коммуникации или сам автор (внутренняя речь, спор самого с собой, разговор с собой вслух, дневник и т.п.).

– Категория всеобщего адресата включает как уже упоминавшийся вариант *всем, всем, всем*!, так и аудиторию радиопередач, зрителей ТВ, читателей печатных изданий, и т.п., то есть всех тех, кто получит данную информацию и ознакомится с ней.

– Недоопределенный адресат, хотя и входит в категорию всеобщего, но выделяется из нее некоторым особым типом восприятия или специальными обстоятельствами КА. В отличие от масскультуры (широкая печать, популярная музыка и т.п.), данная категория включает относительно узкие жанры, рассчитанные на небольшое число способных к адекватному восприятию из числа многих получивших. К этой же категории можно отнести потенциального получателя брошенной в воду бутылки с запиской, адресата обращения: *есть тут врач (кто-нибудь, говорящий по-английски)?* или объявления *ищу блондинку для совместного отпуска.*

3.3.Цель и содержание

3.3.1. В общем случае функция и содержание КА подчинены его цели (задаче). Однако всякий отдельный акт, участвующий в сложном коммуникативном процессе может совмещать некоторые роли из списка *фоновый,*

096-Narinjani:_Layout 1 18.05.2008 16:33 Page 627

отдельный, сложный и *структурный*, о котором мы говорили выше в п.2.2. В этом плане задача отдельного КА может быть предельно локальной (например, передать информацию и выйти из контакта) или подчиненным более сложному КА, частью которого он является. В качестве подчиненного КА может решать различные коммуникативные задачи:

– *Оперативные*: установить \ завершить отдельный контакт или часть более сложного сценария (*этот вопрос мы решили, давайте вернемся к основной теме*).

– *Тактические \ технические* в рамках ведения процесса коммуникации

– *Стратегические* - выполнять основную задачу верхнего структурного КА.

Проиллюстрируем эти уровни на примере: *Извините, как пройти к ближайшему метро?* Если это просьба об информации, то даже в этом разовом КА имеются все три перечисленные составляющие: оперативная - вступить в контакт, тактическая – сформулировать вопрос и стратегическая – получить нужную информацию.

При этом, выполняя разовую задачу, данный КА может открыть локальный диалог, связанный с выяснением нужных деталей, и в любом случае завершающийся оперативным *спасибо!*

3.3.2. В качестве иллюстрации многоуровневости перечисленных коммуникативных задач, рассмотрим их структуру в случае, когда этот вопрос был предлогом для знакомства и данные три составляющие являются компонентами его внешней функции отдельного РА нижнего уровня, в то время как он включен автором в целый комплекс задач более высокого уровня:

- *Оперативно* - начало знакомства:
  - *оперативно*: вступить в контакт,
  - *тактически*: заинтересовать,
  - *стратегически*: познакомиться.
- *Тактически* – закрепление знакомства:
  - *оперативно*: развить диалог общения,
  - *тактически*: закрепить контакт (получить право на его продолжение - следующая встреча),
  - *стратегически*: получить право на его развитие (договориться о возможности развития - приглашения в гости \ в компанию...).
- *стратегически* - окончательное решение:
  - *оперативно*: договориться о конкретном шаге развития,
  - *тактически*: перейти к следующей фазе близости.
  - *стратегически*: добиться близости.

Эта схема полна, если близость и есть конечная цель. Если конечная цель - длительное знакомство, то это схема тактического этапа, и если брак, то может быть и оперативным этапом.

Заключение

Подведем итоги.

Туман уже упоминался в самом первом абзаце, так что для определения жанра этого доклада осталось доуточнить картину post factum. Популярный образ «ежик в тумане» точно передает результат: отдельные кирпичи выглядят до известной степени понятными и даже относительно знакомыми, но от этого ни место каждого кирпича в здании, ни само это целое намного яснее не становится.

В попытке структурировать рассматриваемое пространство, т.е. разрезать туман на части, с риском значительно превысить допустимый объем доклада, были проведены многочисленные сечения для разделения на категории и вычленения основных составляющих, - в сумме более сорока.

Очевидно, что понятия *коммуникация* и *коммуникативный акт* автору сделать обозримыми пока не удалось. В основном данная попытка споткнулась на определении *информации*, для которой в данном контексте никак не удавалось наметить достаточно определенные рамки. Шенон остается прав (см. эпиграф) несмотря истекшие пол века.

Осталось добавить, что список литературы ограничен единственной добавленной к эпиграфу ссылкой, поскольку уже упоминавшийся невероятный объем публикаций делает невозможным включение в такой список даже самой урезанной выборки.

Список литературы

1 Шеннон К. Работы по теории информации и кибернетике. – М.: изд. Иностр. Лит., 1963. – с. 667-668

2 Нариньяни А.С.. Манипуляция как Коммуникативный Акт. // Труды Международной Конференции Диалог’2006 «Компьютерная лингвистика и интеллектуальные технологии». – М.: Изд. РГГУ, 2006. С. 623 – 629.

097-Annotations:Layout 1 18.05.2008 23:04 Page 628

ABSTRACTS

THE CONTEXT SCHEMA OF PREDICATE ARGUMENTS FOR AUTOMATIC EXPANSION OF A DOMAIN ONTOLOGY

Azarova I.V. (ivazarova@gmail.com), Grebenkov A.S. (shurix@grebenkov.ru),Lando T.M. (tatiana.lando@gmail.com)
Saint-Petersburg State University

In the paper the fact mining system Factus is described, it is a prototype model oriented to a restricted domain, which is apt to widen. The problems of domain ontology representation and its extension on the basis of extracted features during text processing are discussed aiming at so called "open concept frame".

The text analysis is accomplished by means of special structures including syntactic arrangements of predicate arguments, their context markers being used for pattern validation.

RUSSIAN AND ENGLISH EMOTIONAL CONCEPTS

Apresjan V.Ju. (valentina.apresjan@gmail.com)
Institute of Russian Language, Moscow

The paper outlines a new method of cross-linguistic comparison of emotion concepts, where entire emotion "clusters" rather than individual terms are juxtaposed. The method is applied to eleven emotion clusters in Russian and English languages. The paper considers both universal semantic tendencies and specific linguistic means in the expression of emotion. The paper proposes certain tentative explanations for the observed cross-linguistic similarities and discrepancies.

ON A PROJECT OF A PRODUCTION DICTIONARY OF RUSSIAN

Apresjan Ju. D. (apr@iitp.ru)
Institute of Russian Language, Moscow

The paper is concerned with a project aimed at creating a production dictionary of contemporary Russian. Work on the project started in 2006 at the Russian Language Institute of RAS. The main idea of the dictionary is to present a complete and unified account of all linguistically relevant properties of each lexical unit. Apart from grammatical forms and senses they include a) regular semantic modifications of the dictionary definition in verifiable contextual conditions, b) detailed government patterns and their possible modifications, c) a list of minor type sentences specific for a given lexical unit, d) its combinatorial potential (especially as handled by the theory of lexical functions), e) its lexicalized prosody. All these make an integral part of the linguistic competence of speakers and should be characterized on the basis of the latest theoretical findings of linguistic research in the respective fields.

REGIONAL VARIANTS OF THE URBAN REALTY TERMS

Akhmetova M.V. (malinxi@rambler.ru)
Journal "Zhivaia Starina" (Moscow, Russia)

The paper deals with the Russian regional terms, describing urban realty — names of different types of apartment houses and flats (depending on the time of building, planning, material, etc.). These words, as a rule, are rarely included into the explanatory dictionaries, except of some colloquial words, which are normative for the speech of Moscow and St.-Petersburg citizens. The research was carried out on the materials of the Integrum database, including mass media publications and public documents from all of the Russian-speaking space. Using the statistics of mentioning these terms in the regional and central public documents, helps to make preliminary conclusion about their areal distribution.

097-Annotations:Layout 1 18.05.2008 23:04 Page 629

AGAINST DECOMPOSITION OF MEANING: RECOGNITION IN SEMANTICS OF IDIOMS

Baranov A.N. (baranov_anatoly@hotmail.com)
Institute of Russian Language, Moscow

In the report the problem of inner form representation in definitions of idioms is discussed. Decomposition of meaning cannot be used for semantic description of non-discrete semantic phenomena such as metaphor and image. It is proposed to use for semantic representation of inner form the strategy of recognizing of metaphor. The process of recognition is supported in a definition of idiom by semantic "trigger", which generates the necessary chain of associations.

LEXICOGRAPHY OF PROVERBS

BelikovV.I. (vibelikov@mtu-net.ru),
Institute of the Russian Language (Russian Academy of Science)

The dictionaries of Russian proverbs are analyzed with respect to their repertory and the selection of the main variant of items.

ORTHOGRAPHY IN THE INTERNET: THE ANALYSIS OF ONE MISSPELL

Bogdanov A.V. (bidon@inbox.ru)
Moscow State University

In the paper we discuss the orthography in the Internet and we analyse a widespread misspell which is writing the soft sign in the ending of verb forms containing suffix *-s'a (-ся)*, like *delaet's'a (делается)*. Our analysis shows that the number of such misspells allows us to talk about kind of new standards in written language of the Internet.

THE CORPUS OF SPOKEN RUSSIAN: DESIGN PRINCIPLES AND APPROACHES TO DATA ANALYSIS

Bogdanova N.V. (nvbogdanova_2005@mail.ru), Brodt I.S. (brodt_05@mail.ru), Kukanova V.V. (vika.kukanova@gmail.com), Pavlova O.V. (olgapavlovaspb@mail.ru), Sapunova E.M. (kaverita@yandex.ru), Philippova N.S. (ninaphilippova@gmail.com)
St. Petersburg State University

The paper reports principles for balancing the corpus of spontaneous monologues in the Russian language collected according to shared linguistic and sociolinguistic parameters. It presents samples of collected data, benefits of multilevel analysis and perspectives of further augmentation.

"JA NE BYL…MENJA NE BYLO.." OR HOW MANY DIFFERENT BYT' (BE) IN RUSSIAN

Borschev V.B. borschev@linguist.umass.edu
VINITI RAS & UMass

This work introduces and analyzes the Russian example Ja ne byl v zale, kogda vyklučili svet 'I wasn't in the hall when they turned out the lights.' This example refutes Ju.D. Apresjan's claim that sentences of that kind cannot have a "synchronous" interpretation. Various meanings of the verb byt' 'be' in locative and existential sentences are discussed.

COMPARISON OF FIVE METHODS FOR VARIABLE LENGTH TERM EXTRACTION

Braslasvki P.I. (pb@imach.uran.ru), Sokolov E.A. (esokolov@list.ru)
Institute of Engineering Science UD RAS, Ekaterinburg

The paper investigates and compares five methods for variable length term extraction and assembling. Experiments are conducted on a corpus of scientific papers on genetics and microbiology. Evaluation method combining both expert and formal assessment is proposed, the results of comparative evaluation of the methods are presented.

097-Annotations:Layout 1 18.05.2008 23:04 Page 630

MULTITASKING SEARCH: FACT ARTIFACT, NEGLIGIBLE EXCEPTION?

Buzikashvili N.E. (buzik@cs.isa.ru)
Institute of System Analysis, Russian Academy of Sciences

The paper considers search on the Web. Questions on the users' manners of search are formulated, with emphasis on multiple tasks execution. It is shown that multitasking is rare, usually includes only two task sessions and is formed into a temporal inclusion of an interrupting task into the interrupted one. Quantitative characteristics of search behavior in 3 classes of temporal sessions (single-task session, several tasks executed one-by-one, and multitasking session) were compared, and significant differences were revealed.

COMPLEX TECHNOLOGY OF AUTOMATIC TEXT CLASSIFICATION

Vasilev V.G. (vvg_2000@mail.ru)
Institute of Informatics Problems of the Russian Academy of Sciences (IPI RAN)

The report discusses the problems that arise when building automatic text classification systems. Main elements of the integrated text classification technology are described. Particular attention is given to the construction of combined decision rules for the implementation of a hierarchical classification of texts.

COMPARATIVE LEXICOGRAPHIC DESCRIPTION OF RUSSIAN WORDS AND GESTURES OF RUSSIAN SIGN LANGUAGE IN RUSLED DICTIONARY

Voskresenskiy A.L. (AVoskresenskij@college.mesi.ru)
College MESI

Comparative lexicographic description in RuSLED dictionary of Russian words and gestures of Russian sign language with same or near meanings is presented. There is intended to use multimedia dictionary RuSLED for study of Russian words and gestures of Russian sign language usage features.

GENETIC ALGORITHM FOR AUTOMATIC DIVISION OF WORS INTO MORPHEMES

Gelbukh A.[1] (www.gelbukh.com), Sidorov G.[1] (www.cic.ipn.m/~sidorov), Lara-Reyes D.[1],
Chanona-Hernandez L.[2], Chubukova M.[3] (licht66@mail.ru)
[1] Natural Language and Text Processing Laboratory,
Center for Research in Computer Science, National Polytechnic Institute,
Av. Juan Dios Batiz, s/n, Zacatenco, 07738, Mexico City, Mexico
[2] ESIME Zacatenco, National Polytechnic Institute,
Zacatenco, 07738, Mexico City, Mexico
[3] Philological education department,
Moscow Institute of Continuous Education, Moscow, Russia

The paper discusses unsupervised technique for automatic detection of morpheme structure of words in flexive languages, using Spanish language as a case study. We use global optimization implemented as genetic algorithm, without any heuristics or assumptions that affect the problem dimensions *a priori*. Description of genetic algorithm is given; preliminary results of evaluations are presented. Input data is the list of words, compiled on the basis of a dictionary or a corpus. Output data is the same list of the words separated in morphemes. As many other automatic methods, this algorithm does not pretend to detect a hundred percent correct results and require postprocessing. Still, it allows for fast detection of tendencies in data and for obtaining of preliminary results without manual work.

TWO MEANINGS, TWO LINGUISTIC ITEMS? RUSSIAN „AHA" IN SPONTANEOUS DIALOGUE

Gerassimenko O. (olga.gerassimenko@ut.ee)
University of Tartu, Estonia

Russian dialogue particle „aha" can express either agreement/confirmation or surprise/satisfaction. In Russian lexicography those meanings are mostly presented as being expressed with homonymic linguistical items, the particle and the interjection. The paper examines examples of „aha" in spontaneous institutional dialogues and discusses the possibility of finding a common meaning part.

097-Annotations:Layout 1 18.05.2008 23:04 Page 631

MACHINE TRANSLATION SYSTEM TILDE TRANSLATOR: NEW STAGE OF ENGLISH / RUSSIAN-LATVIAN MACHINE TRANSLATION

Gornostay T. (tatjana.gornostaja@tilde.lv)
company Tilde, Riga (www.tilde.com)

The article presents a new linguistic software – machine translation system Tilde Translator, and focuses on lexical ambiguaty and multiword expression treatment in the system. The improvement of machine translation quality with semantic filters in the system is also described.

RUSSIAN ADJECTIVE PHRASE: SPLIT AP HYPOTHESIS

Grashchenkova A.E. (izmaja@mail.ru)
Russian State University for Humanities

The paper presents a minimalist approach to Russian Adjective Phrase (AP hereafter) structure. The puzzling properties of predicative long form (LoF) adjectives with complements is the starting point of the paper.

To explain the distribution of the complement-taking adjectives, we suggest the multi-layered structure of adjectival phrase. The internal A is a lexical head that surface as a short form (ShF) adjective. External small a is a functional head responsible for case concord of attributive LoFs.

The chief claim concerns the properties of lexical A heads in Russian. These ShF phrases: (i) are the locus for argument merging; (ii) project their own Spec position; (iii) do not assign structural case (iv) allow eventive (stage-level) interpretation. At the same time, the external LoF a-shell lacks all these properties and is responsible for case-concord in noun phrases.

As for the constraint on complementation, attested with Russian predicative LoF adjectives, we supposed that it is due to the two facts: "defective" structure of nominative predicates on the one hand and the elaborated shell structure of adjectival phrase on the other. In such constructions the subject of the lexical AP "has not enough time" to raise to Spec, IP and activate case feature on I, which subsequently should be transmitted to the a head through Pred. This conflict does not arise in case of instrumentals (assigned by Pred) and ShF (no case assignment).

Then, case features on LoF do not influence its complement-taking potential in attributive function and in secondary predication. We ascribe the grammaticality of attributive and secondary instrumental / nominative LoFs on the fact that such adjectival phrases are control structures and the case value does not dependend on the internal subject raising.

The proposed analysis is supported by several other properties of LoF and ShF: distribution of symmetric predicates, stage/individual-level interpretations, properties of derived nominals and others.

CORPUS OF ORAL RUSSIAN IN THE FRAMEWORK OF RUSSIAN NATIONAL CORPUS. CONSTRUCTION PROJECT

Grishina E.A. (rudi2007@yandex.ru), Savchuk S.O. (savsvetlana@mail.ru)
Institute of Russian Language,Moscow

The paper describes the construction project "Corpus of Oral Russian", which may be created on the basis of the Movie Sub-Corpus of the Russian National Corpus. The authors offer some solutions to the problems concerning the structure of the Corpus, the types of the annotation, the format of the issues, the types of the queries, and the variety of the tasks which may be posed and solved by the use of the Corpus.

DESCRIBING SHAPE: INSTRUMENTAL CONSTRUCTION «X Y-OM»

Dessiatova A.V. (patine@gmail.com), Russian State University for Humanities
Lashevskaja O.N. (olesar@mail.ru), V.V.Vinogradov Institute of the Russian Language
Mahova A.A. (discourse@yandex.ru), Moscow State University

The paper analyzes the semantics of Russian instrumental construction with the meaning of shape (*xvost kol'com* 'ring tail', slozhit' gubki bantikom *'to make Cupid's bow*'). Spatial interpretation of this construction is described in terms of topological classes (Talmy 2000, Rakhilina 2000). Possible mirroring of topological classes in both slots X and Y is investigated as well as their predictable mutual accommodation.

097-Annotations:Layout 1 18.05.2008 23:04 Page 632

DEIXIS WITHOUT SPEAKER: TOWARDS THE SEMANTICS OF THE GERMAN DEICTIC ELEMENTS *HIN* AND *HER*

Dobrovol'skij D.O. (dm-dbrv@yandex.ru), Russian Academy of Sciences, Russian Language Institute
Padučeva E.V. (elena708@gmail.com), Russian Academy of Sciences, Institute of Scientific and Technical Information (VINITI)

Semantics of deictic words can be analysed more efficiently if we take the communicative situation of the utterance into account. Traditionally, the semantics of the German deictic elements hin and her was described as being orientated towards the speaker – towards the speaker's place and time. However, this is true only for the canonical communicative situation, when speaker and hearer are both in the same place. In non-canonical situations (when speaker and hearer are not in the same place) and especially in contexts of hypotaxis or narrative, the speaker may be deprived of his "deictic privileges", which are then transferred to some other persons.

THE PARSER OF ETAP-3 LINGUISTIC PROCESSOR: EXPERIMENTS ON RANKING SYNTACTIC HYPOTHESES

Druzhkin K.Ju. (druzhkin@iitp.ru), Tsinman L.L.. (cinman@iitp.ru)
Institute for Information Transmission Problems

An attempt is made to optimize the operation of the parser in ETAP-3 linguistic processor. The idea is to change parsing rules in such a way that the emerging syntactic hypotheses be ranked according to probabilities of their appearance in the resulting syntactic tree of the sentence processed. Experimental results are given.

AUTOMATIZATION OF AN ONTOLOGICAL ENGINEERING FOR SYSTEMS OF KNOWLEDGE MINING IN TEXT

Ermakov A.E. (ermakov@metric.ru)
RCO, Moscow

The present report is devoted to the problems of using ontologies in text mining systems. Peculiarities of ontologies used in such systems are examined. A method for automatic ontology generation, in which terms of data domain and relations between them are initially detected by means of computer analysis of the text, is proposed.

A.PLATONOV'S TEXTS AS A LINGUISTIC SOURCE

Zalizniak Anna A. (anna-zalizniak@mtu-net.ru)
Institute of linguistics, Russian Academy of Sciences

Anomalous phrases in A.Platonov's texts so far have been investigated exclusively as a source of information on the author's poetic world. The paper demonstrates that Platonov's linguistic anomalies can be used as a source of information about Russian language. These anomalies reveal some subtle semantic, combinatorial and categorical properties of Russian words, which hardly could have been noticed otherwise. This information can be used in explanatory dictionaries of Russian, as well as in the semantic tagging of electronic corpora.

TERMS FOR SCIENTIFIC AND TECHNICAL KNOWLEDGE REPRESENTATION IN DIGITAL SPHERE

Zatsman I.M. (im@a170.ipi.ac.ru), Kurchavova O.A. (koa@a170.ipi.ac.ru)
Institute for informatics problems of the RAS

Documents of the 7-th Framework program of the European Union, accepted for the period 2007-2013, contain formulations of the new tasks concerning to the knowledge representation problem in the digital sphere. In the paper key positions of these formulations are analyzed. Results of the analysis are used for definition of some terms suggested for the description of knowledge representation processes in digital libraries.

097-Annotations:Layout 1 18.05.2008 23:04 Page 633

THE IDEA OF MATCHING NAMES IN RUSSIAN

Iomdin B.L. (iomdin@ruslang.ru)
Russian Language Institute

The paper deals with metalanguage lexical units that convey certain relations of names of different objects: these are Russian units одноимённый 'of the same name, cognominal' (and its derivates) and *так и называется* » 'called exactly this way'. Such items are difficult to interpret in NLP applications. Lexicographic definitions are proposed based on a number of key senses identified by the author: ideas of coincidence, correspondence, and simplicity.

IN THE DEPTHS OF MICROSYNTAX: A LEXICAL CLASS OF SYNTACTIC IDIOMS

Iomdin Leonid L. (iomdin@iitp.ru)
Institute for Information Transmission Problems, Russian Academy of Sciences

A class of Russian syntactic idioms is considered from the theoretical and NLP points of view. The class, formed with the noun сила 'force, power' consists of a variety of lexical units with surprisingly individual peculiarities. Examples of this class include (1) a preposition в силу ≈'by virtue of', as in В силу этой теории поведение в одной точке вселенной влияет на поведение в другой точке 'By virtue of this theory, the behaviour in one point of the universe influences the behaviour in another point'; (2) an adverb of degree от силы 'at the most', as in от силы десять человек 'ten people at the most', (3) an adverbial pattern в X-овую силу 'using such and such part of one's force, as in работает в полную силу 'he works to the full extent of his power', работает в треть силы 'He works using a third of his force'; (4) a predicative adverb в силах 1 ≈ 'being able' as in старик был не в силах быстро ходить 'the old man was unable to walk fast', (5) a predicative adverbial pattern в (чьих-либо) силах 2 'within one's power', as in сдержать смех было не в моих силах 'to contain laughter was beyond my powers'. Specific descriptions of several of these idioms are given using a specially designed standard layout.

SPEAKER'S PROSODIC PORTRAIT AS A TOOL OF SPOKEN DISCOURSE TRANSCRIPTION

Kibrik A.A. (kibrik@comtv.ru)
Institute of linguistics, Russian Academy of sciences

A methodological tool is proposed that enhances the quality of discourse transcription, in the course of preparing corpora of spoken language. Prosodic prototypes unerlying discourse segmentation and expression of phasal meanings can be identified with the help of prosodic portraits of individual speakers.

BUILDING A GRAPH OF SEGMENT CONNECTIONS (RUSSIAN SENTENCE SURFACE-SYNTACTICAL ANALYSIS)

Kobzareva T.Yu. (stamstam@mtu-net.ru)
Russian State University for Humanities

The paper discusses a linguistic basis of the segment connections building in Russian sentence and some problems that arise when searching for a control word of a specified segment.

UNICELLULAR ORGANISMS OF COMMUNICATION UNDER A MICROSCOPE: GERMAN PARTICLE *JA* VERSUS ITS RUSSIAN TRANSLATION EQUIVALENTS *VED'* AND *ŽE*

Kobozeva I.M. (kobozeva@list.ru), Orlova S.V. (svetlachok-star@yandex.ru)
Lomonosov Moscow State University

In the paper German modal particle ja in constative utterances is compared to its Russian translation equivalents VED' and ŽE on the basis of studying parallel samples of modern German prose and its professional translations into Russian. The analysis reveals the following differences:

VED' presupposes its proposition as a fact while ja and ŽE do not, and it explains the ability of the latter two to be freely used in imperatives;

097-Annotations:Layout 1 18.05.2008 23:04 Page 634

VED' and ŽE specify the degree of rhetoric activity (≈ intencity of illocutionary force) as normal and high resp. while for JA this semantic feature is irrelevant and this makes the choice of its translation equivalent dependant on such pragmatic features of the context as its relation to speaker's interests and interpersonal relations among the interlocutors;

JA can be used in responses, implying yes / no answers to direct questions, while VED' and ŽE cannot occur in this context;

the use of VED' in answers demands the dictal component of its propositional content to be different from that of the question;

VED' cannot occur in correcting remarks and direct answers if it is not preceded by initial adversative particles (NO, A, DA). Its use together with one of these particles overtly marks the response as conflicting with some of the addressee's initial assumptions and thus violating the maxim of consent and so in some cases it may damage semantic equivalence of the translation with respect to the interpersonal aspects of utterance meaning.

DATABASE «INTONATION OF RUSSIAN INFORMATIONAL TEXTS»

Kodzasov S.V. (sankod@philol.msu.ru), Arkhipov A.V. (arxipov@philol.msu.ru), Zakharov D.M. (leon@philol.msu.ru), Krivnova O.F. (okri@philol.msu.ru)
Moscow State University

The development of a data base for intonation of oral mass-media texts is now in progress at the Philological Faculty of Moscow University. A highly detailed system for sentence prosody description is used. Great differences are found between the use of prosodic means in informal dialogues and in informational texts in TV-programs.

CLASSIFICATION SCHEME OF THE SEMANTIC DICTIONARY OF THE MONITORING SYSTEM: TEST APPLICATION TO EVALUATION OF SCIENTIFIC WORK' PERFORMANCE

Kozhunova O.S. (kozhunovka@mail.ru, okozhunova@ipiran.ru)
Institute for Informatics Problems of the Russian Academy of Sciences

A brief description of the experiment on the evaluation of the scientific work' performance in Russian Academy of sciences carried out in 2007 is given. At the final stage this action revealed several problems. In this connection, a method and an instrument of their solution are suggested. These are classification method and semantic dictionary with integrated classification scheme, correspondingly.

'CAUSE' OR 'ENABLE': ANALYSIS OF CAUSATIVE VERBS SEMANTICS

Kozlova A.V. (avkozlova@rambler.ru), Lutikova E.A. (katjal@philol.msu.ru), Fedorova O.V. (olga.fedorova@msu.ru)
Lomonosov Moscow State University

In this paper, data of the experimental investigation of Russian causative verbs semantics is presented. The investigation was conducted in the framework of Force dynamics theory. We distinguish the concepts of CAUSE, ENABLE, and PREVENT depending on the correlation of three main parameters of the causative situation: 1)the tendency of the patient for a result, 2) the presence of opposition between the affector and the patient, and 3) the occurrence of a result.

EVOLUTION OF ARGUMENTATIVE LANGUAGE BEHAVIOUR PATTERNS AS AN ASPECT OF COMMUNICATIVE COMPETENCE GENESIS

Kolmogorova A.V. (nastiakol@mail.ru)
Kouzbass state pedagogical academy

The article deals with the study of language behavior patterns using by Russians in the communicative situation of argumentation.

PAUSES ON THE DIFFERENT TYPES OF SYNTACTIC BOUNDARIES IN JAPANESE: A CORPUS STUDY

Komarova A.D. (komarovochka@gmail.com)
Russian State University for Humanities

The present research is concerned with the pauses at different syntactic boundaries in oral monologue Japanese speech. It aims to find out, how frequent and therefore probable are the pauses at the boundaries of sentences and clauses lesser than sentences and what their "normal" length is.

097-Annotations:Layout 1 18.05.2008 23:04 Page 635

PROSODY OF CLAUSE-COMBINING IN RUSSIAN: A CORPUS-BASED CASE-STUDY

Korotaev N.A. (n_korotaev@hotmail.com), Podlesskaya V.I. (podlesskaya@ocrus.ru)
Russian State University for Humanities (RSUH)

The paper reports a corpus-based study of prosodic strategies employed in multiclausal structures with a postpositioned dependent clause in spoken Russian. Three main strategies are discussed: (1) the pitch direction at the primary accent in the main clause is opposite to that in the dependent clause, (2) the pitch direction at the primary accent in the main clause copies that in the dependent clause, and (3) the main clause remains non-accented. Quantitative and qualitative analysis is provided to explain the speaker's choice between the three strategies.

CONTROLLING DYNAMIC SPEECH BEHAVIOUR OF VIRTUAL COMPUTER AGENTS

Kotov A.A. (kotov@harpia.ru)
Russian State University for Humanities

We represent and discuss a model to control speech behaviour of a virtual computer agent (computer game agent, interface component or, in the future, mobile robot). The model simulates "mood dynamics", which controls agent's behaviour in a communication. In particular, the model uses a set of phrasal templates to construct shorts monologues, revealing the dynamics of agent's "feelings" and allowing the agent to switch between several dialogues in a communication.

MECHANISMS OF INTERATION BETWEEN VERBAL AND NOHVERBAL UNITS IN A DIALOG II B. DEICTIC GESTURES AND SPEECH ACTS

Kreydlin G.E. (gekr@iitp.ru) Russian State University for Humanities

Academic lecture regarded as a kind of a dialog is a suitable experimental ground for studying general regularities and specific rules of gesture-speech interrelation and human interaction. In the first part (part II A) of the research a classification of didactic deictic gestures has been compiled and some classes of these gestures has been described. In this part (part II B) of the research I imply to demonstrate that deictic gestures of each type have their own non-trivial relations with the verbal and nonverbal signs in a dialog.

EVALUATING OF FREQUENCY OF SYNTACTIC MOLECULES (ON THE EVIDENCE FROM THE RUSSIAN GENERAL CORPUS)

Krylov Sergej A. (krylov-58@mail.ru)
Institute of Oriental Studies, Russian Academy of Sciences, Moscow, Russia
Institute of Systemic Analysis, Russian Academy of Sciences, Moscow, Russia

An attempt is made, to evaluate the frequency of syntactic molecules (= minimal autosemantic sentence parts, able to serve as answers to a question) on the evidence from the Russian General Corpus (created on the base of the Uppsala Corpus) with the help of the StarLing database processing software package.

БЛАГОРОДНЫЙ: LANGUAGE CONCEPTION OF CONNECTION BETWEEN INTERNAL QUALITIES AND BIRTH OF PERSON

Krylova T.V. (ta-kr@yandex.ru)
Institute of Russian Language of Russian Science Academy

The objects of this article are words *благородный* and *великодушный*. Firstly, we describe the difference in their semantics and try to establish the connection between meaning of благородный and its internal form. Then, the polysemy of *благородный* is examined. At last, we analyse the meaning of lexemes *благородный* 3.1 и 3.2 (*благородное лицо, благородное животное*) and formulate the hypothesis that the conception of connection between internal qualities and birth of person is still preserved in modern language.

097-Annotations:Layout 1 18.05.2008 23:04 Page 636

TEXTUAL DIALECT CORPUS AS A MODEL OF TRADITIONAL RURAL COMMUNICATION

Kryuchkova O.I. (vpks@rambler.ru), Goldin V.E.(goldinve@yandex.ru)
Saratov State University N.G. Chernyshevkij

The report deals with the principles of organization and methods of building a multimedia textual dialect corpus, representing dialect as a comprehensive whole of cultural and communicative features and modeling the communication of specific speech groups in specific social and cultural environment.

USEFUL EXTENSIONS TO TRANSLATION-ORIENTED TERMINOLOGICAL DICTIONARIES (CASE: TWO FINNISH-RUSSIAN DICTIONARIES)

Kudashev I.S. (igor.kudashev@helsinki.fi), Kudasheva I.O. (irina.kudasheva@helsinki.fi)
University of Helsinki

In this article, we describe some useful extensions to translation-oriented terminological dictionaries using as an example two dictionaries compiled at the University of Helsinki, Palmenia Centre for Continuing Education in Kouvola, in 2003–2007. These dictionaries are mostly descriptive but they contain some elements which are usually characteristic of normative dictionaries, such as restrictive labels, strict terminological definitions, and concept charts. Special attention is paid to translator-friendly techniques, such as explicit marking of partial and artificial equivalents and explanation of the differences between concepts in the source and target languages.

LINGUISTIC PROCESSOR “SEMANTIX” FOR KNOWLEDGE EXTRACTION FROM NATURAL TEXTS IN RUSSIA AND ENGLISH

Kuznetsov I.P.(igor-kuz@mtu-net.ru),
Institute for informatics problems of the RAS
Efimov D.A. (d.efimov@synsys.ru),
Synergetics Systems

Paper considers the linguistic processor ”Semantix” for automatic formalization of natural language texts in some fields: criminal, autobiography, texts about terrorism . The processor extract from texts the user objects, their links and facts of object actions. Results are XML-files which are used for Knowledge Base organization, semantic search and analytic tasks.

PARTS OF SPEECH SYNCRETISM IN RUSSIAN AND URALIC (THE PROBLEM OF THE COMPOSITION OF BILINGUAL DICTIONARIES FOR INFLECTIONALAND AGGLUTINATIVE LANGUAGES)

Kuznecova A.I. (aikuznec@yandex.ru)
Moscow State University

Parts of speech syncretism in Russian and Uralic (the problem of the composition of bilingual dictionaries for inflectional and agglutinative languages)

The article discusses pro and contra of presenting the material as homonyms or polysemantic units in bilingual dictionaries.

ABOUT «NON-NOMINATIVE» DICTIONARIES (LEXICAL DATABASES)

Kustova G.I. (galina03@mtu-net.ru),
Moscow State Pedagogical University

This report deals with a project of dictionary (lexical database) including «non-nominative» items which are used as adverbial modifiers (for ex. на ходу, под предлогом (чего), во всяком случае).

097-Annotations:Layout 1 18.05.2008 23:04 Page 637

WEB-SPACE AND MATERIALS OF NEWS AGENCIES

Lande, D.V. (dwl@visti.net), Brajchevskiy, S.M.(smb@visti.net), Darmokhval, A.T. (hval@visti.net), Morozov, A.Y., (alex@visti.net)
ElVisti Information Center, Ukraine, Kiev

In this article we investigate to what extent materials available to paying subscribers are openly published on web-sites. We obtained the distribution of news agencies' messages based on the time of delay. We also measured specific quantity of reprints of the news agencies' materials on web sites as well as Internet messages included to the agencies' news-lines.

THE RIDDLES OF THE RUSSIAN PARTICLE UZH

Levontina I.B. (irina.levontina@mail.ru)
Institute of Russian Language, Moscow

Russian discourse particle *uzh* is very difficult to describe. It produces manifold pragmatic effects, and it is unclear how this effects are connected with the components of its meaning. The paper is devoted to some of such components and discourse effects they cause.

U NAS U STATJI NAZVANIE NE PRIDUMALOS': RUSSIAN CONSTRUCTIONS WITH RECURSIVE NP WITH PREPOSITION U

Leonteva A. L. (njusha-nn@mail.ru),
Moscow State University
Leontev A. P. (taonick@yandex.ru)
Moscow State University, ABBYY

Construction with two NPs with prepositions u, such as U men'a u dočki segodn'a den' roždenija 'Today is my daughter's birthday' – is very often used in colloquial Russian. In this paper I will describe syntactic, semantic and discourse features of this construction.

ASYNDETON AND COLON. TRANSCRIBING SPOKEN NARRATIVE

Litvinenko A.O. (allal1978@rambler.ru)
Moscow State University

The paper is devoted to a closed class of Russian asyndetic composite sentences that require the use of colon in written language and are characterized by a special intonation in spoken language. The problems that arise while transcribing such sentences in spoken narrative are discussed.

AN ALGORITHM OF TEXT SEGMENTATION ON SYNTACTIC SYNTAGMAS FOR TTS SYNTHESIS

Lobanov B.M. (lobanov@newman.bas-net.by)
United Institute of Informatics Problems, National Academy of Science of Belarus

An algorithm of segmentation of the text on the syntactic syntagmas, based on the analysis of the steady phraselogical and grammar-semantic word-combinations making the sentence is suggested. The basic sense of allocation consists in the sentence of considered word-combinations that now freedom of its division into syntagmas is limited, namely: the syntagma border can be only outside of word-combinations, but not in them.

«INTOCLONATOR» - A COMPUTER SYSTEM FOR PROSODIC SPEECH PARAMETERS CLONING

Lobanov B.M. (Lobanov@newman.bas-net.by), Tsirulnik L.I. (L.Tsirulnik@newman.bas-net.by), Sizonov O.G. (Osizonov@yahoo.co.uk)
United Institute of Informatics Problems on National Academy of Science of the Republic of Belarus

A computer system of prosodic speech parameters cloning is described. The system allows to automate the process of creation of a complex prosodic portraits necessary for TTS synthesis. The system is intended for widening of inventory of prosodic portraits for the personalized speech synthesis under texts of various genres.

097-Annotations:Layout 1 18.05.2008 23:04 Page 638

AUTOMATED ANALYSIS OF MULTIWORD EXPRESSIONS FOR COMPUTATIONAL DICTIONARIES

Loukachevitch N. (louk@mail.cir.ru), Dobrov B. (dobroff@mai.cir.ru), Chuyko D. (Dasha_C@mail.ru)
Research Computer Center Moscow State University (MSU NIVC);
NCO Center for Information Research

In the paper we describe the development of an automatized system for analysis of multiword expressions that facilitates the discovery of specific features of syntactic and semantic behaviour of multiword expressions. The analysis is based on automatic comparison of the component structure of expressions and uses the knowledge described in a thesaurus-like lingustic resource. At present we test the system in the process of terms acquisition for Ontology on natural sciences and technologies.

FREQUENCY DICTIONARY OF THE RUSSIAN NATIONAL CORPUS: PRINCIPLES AND TECHNOLOGY

Lashevskaja O.N. (olesar@mail.ru)
Institute of Russian language, Moscow
Sharoff S.A. (s.sharoff@leeds.ac.uk)
University of Leeds, United Kingdom

A frequency dictionary represents the base lexicon of contemporary Russian (1950–2005) that gives information about word frequency in actual use and provides frequency comparisons between different functional styles and periods of creation of texts. The dictionary is based on texts of the Russian National Corpus Словарь (100 million words).

RHETORICAL ENANTIOSEMY IN THE SPEECH CORPUS OF THE RUSSIAN EVERYDAY COMMUNICATION “ONE SPEAKER’S DAY”

Markasova E.V. (markasovaelena@yandex.ru)
St.Petersburg State University

This Report is devoted to the rhetorical antonymy recognized only in oral speech. Unlike inherent and adherent antonymy, the kind distinguished on the paper is characterized not by the inner antinomy of meanings, but by the opposition of the communicative purposes (contructive and destructive).

SYNTAX OF CORRELATIVE CONSTRUCTIONS IN RUSSIAN: A GENERATIVE APPROACH

Olga V. Mitrenina (mitrenina@gmail.com)
State University of St.-Petersburg

Barriers between the correlative clause and the main clause in correlative constructions in Russian are described. It is also shown that correlatives do not reconstruct in Russian. The preliminary syntactic structure of Russian correlatives is suggested, that involves the position of topic and/or focus.

CORPUS ANALYSIS OF SELECTIONAL PREFERENCES OF FREQUENT WORDS IN RUSSIAN

Mitrofanova O.A. (alkonost-om@yandex.ru), Belik V.V. (ogibbion14@pisem.net), Kadina V.V. (veraiii@yandex.ru)
Saint-Petersburg State University

The paper presents results of a corpus-based study of selectional preferences of frequent Russian lexemes. Research procedure requires analysis of co-occurrence data obtained from Russian texts. It is implied that selectional preferences of a lexical item may be defined through sorting its left/right neighbours in bigrams by MI-score values. Given an ordered set of neighbours for a lexical item, it is possible to induce its context patterns. Selectional preferences are specified with respect to morphological and semantic features of co-occurring lexical items.

STATISTICAL WORD SENSE DISAMBIGUATION IN CONTEXTS FOR NAMES OF PHYSICAL OBJECTS

Mitrofanova O.A. (alkonost-om@yandex.ru), Panicheva P.V. (ppolin@yandex.ru), Saint-Petersburg State University;
Lashevskaja O.N. (olesar@mail.ru) Institute of Russian Language, Moscow

The paper presents experimental results on automatic word sense disambiguation. Contexts for Russian nouns denoting physical objects extracted from the National Corpus of the Russian Language serve as an empirical basis of the

097-Annotations:Layout 1 18.05.2008 23:04 Page 639

study. Optimal conditions for WSD are defined taking into account lexical markers of word meanings in contexts and semantic annotation of contexts.

THE INTERPRETING CORPUS AS A NEW TYPE OF TEXT CORPUS

Mikhailov M.N. (mihail.mihailov@uta.fi), Isolahti N.B. (nina.isolahti@uta.fi)
University of Tampere, Tampere, Finland

The issues discussed are the principles of compiling of interpreting corpora with a corpus of court interpreting as an example. Such a corpus combines a spoken corpus with a parallel corpus. The tagging should reflect communicative, prosodic, as well as extralinguistic information. The interpreting corpora are a valuable resource of data for multidisciplinary research.

AN ATTEMPT TO COUNT POETIC DEVICES IN EXCERPTS FROM DIFFERENT WORKS OF V.NABOKOV AND A.PLATONOV

Mikhail Mikheev (m-miheev@rambler.ru)
Moscow State University

Different instances of poet devices used in the works of V.Nabokov and A.Platonov were counted with a list of about 100 pages for each author. Based on this comparison allows to hypothesize an opposition between their respective poetics as the poetic of language elegance and the poetic of awkwardness.

ON THE UNIVERSALITY OF NOUN-VERB DISTINCTION

Mikhina S.M. (sofia_mikhina@mail.ru)
Russian State University for Humanities

The paper is devoted to the typology of languages which show weak noun-verb distinction. I propose that there exist two distinct types of languages within this group: one of them allows verbal and nominal categories to be combined in a phrase, the other does not. In this paper I discuss some properties of the languages of both types, mostly concerned to restrictions on verbs used as arguments.

ON THE DICTIONARY OF CHANGES IN RUSSIAN LANGUAGE GOVERNMENT

Muravenko E. V. (emuravenko@yandex.ru)
Russian State University for the Humanities

The report lays the foundation for the need to compile a new specialized dictionary, reflecting changes in Russian language government over the period from early 19th century to the present day. The author presents a concise list of principles underlying such a dictionary and introduces a sample dictionary article for the verb skuchat'.

THE PRAGUE DEPENDENCY TREEBANK

Nedoluzhko A. (nedoluzko@ufal.mff.cuni.cz), Hajič J. (hajic@ufal.mff.cuni.cz)& Co.
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

The Prague Dependency Treebank 2.0 (PDT 2.0) contains a large amount of Czech texts with complex and interlinked morphological (2 million words), syntactic (1.5 MW) and complex semantic annotation (0.8 MW); in addition, certain properties of sentence information structure and coreference relations are annotated at the semantic level. PDT 2.0 is based on the long-standing Praguian linguistic tradition, adapted for the current Computational Linguistics research needs. The corpus itself uses the latest annotation technology.

Besides the large corpus of Czech, a corpus of Czech-English parallel resources (The Prague Czech-English Dependency Treebank) is being developed. English sentences from the Wall Street Journal and their translations into Czech are being annotated in the same way as in PDT 2.0. This corpus is suitable for experiments in machine translation, with a special emphasis on dependency-based (structural) translation.

In the report, the basic annotation scheme is represented, with special reference to complex semantic (tectogrammatical) level. The system of syntactic functors and valency lexicon VALLEX are also discussed.

097-Annotations:Layout 1 18.05.2008 23:04 Page 640

THE MEANING OF PROSODIC INFORMATION IN LEXIGORAFIC REPRESENTATION OF POLYSEMY AND HOMONIMY

Pavlova A.V. (anna.pavlova@gmx.de)
SAP AG, Walldorf, Deutschland

The lexical semantics of the word can determine its weak or strong accentual position in the phrase, its intention to play the role of the topic or the comment. The bonds between the lexical meaning of the word and its potential accentuality could help to describe the different meanings of one and the same polisemic word in more detail. The interaction between the polisemic word and its accentaulity allows to find its additional specific and particular meanings. The subjective and estimating, negative and retrospective semantics is especially „appealing" for the phrase accent. But there are several factors which can withstand this accent „appeal", for example specific communication task (pure narrativity, explanation of cause, imperative sentences), idiomatic phrases, innumeration, the use of numerals. If we also include the prosodic information about accentuality into the dictionary, it is necessary to comment at least on the potential obstacles which can destroy the anticipated accentual construction of the phrase. This comment could be presented for instance in the foreword of such a dictionary. Generally not all the words of the vocabulary request this kind of prosodic information. On the other hand, there are some lexical meanings of the polisemic words closely connected with the accentual emphasis, this fact should not be neglected in the lexicography.

REGISTER OF INTERPRETATION AS DISAMBIGUATING CONTEXT

Padučeva E.V. (elena708@gmail.com)
Russian Academy of Sciences, Institute of Scientific and Technical Information (VINITI)

The focus of attention in modern semantics gradually transfers from the meaning of separate linguistic entities to meaning shifts and contexts that motivate these meaning shifts. The type of communicative situation (and registers of interpretation it engenders – such as dialogical register, narrative, hypotaxis) is one of the most relevant parameters. Examples are given of egocentrical grammatical categories, words and constructions that have different interpretations in different registers.

INTONATION OF THE GERMAN COHERENT DISCOURSE IN CONTRAST TO THE RUSSIAN ONE

Paljko M.L. (m_palko@mail.ru)
Institute of Linguistics (Russian Academy of Sciences)

It is widely recognized that the marker of text incompleteness in many languages is the rising tone. This paper argues that in German the intonational strategies of the coherence maintaining can be specified and that a variety of ways to show that a statement is not text-final can be singled out.

CORPOREALITY AND PECULIARITIES OF SEMIOTIC BEHAVIOUR IN DIALOGUE

Pereverzeva S. I. (P_Sveta@hotmail.com), Kreydlin G.E. (gekr@iitp.ru)
Russian State University for the Humanities

This paper discusses modification of some syntactic rules that regulate the interaction of verbal and nonverbal semiotic codes in the dialogue. We show that there is regular correspondence between particular meanings in the semantic explanation of the gesture given and different components of the physical realization of this gesture.

ALIGNMENT OF UN-ANNOTATED PARALLEL CORPORA

Kedrova G.E. (kedr@philol.msu.ru), Potemkin S.B. (potemkin@philol.msu.ru)
Moscow State University

Aligning parallel texts, i.e. automatically setting the sentences or words in one text into correspondence with their equivalents in a translation, is a very useful preprocessing step for a range of applications, including but not limited to machine translation, cross-language information retrieval, and dictionary creation. We are presenting a new alignment algorithm for aligning bilingual, linguistically un-annotated parallel corpora. It enables alignment at sentence level, using bilingual dictionary and heuristic cues, along with linguistics-based rules. The program based on the algorithm currently aligns Russian and English texts, requires no previous marking-up or other manual text pre-processing. Russian lemmas are retrieved in the grammar dictionary. The adaptive nature of the system allows experiments with a variety of fic-

097-Annotations:Layout 1 18.05.2008 23:04 Page 641

tion or non-fiction (i.e. scientific and juridical) texts. The algorithm deals with the typical alignment problems like the correct alignment of one-to many sentences correspondence and omission of a sentence, or how to align texts with different syntactic patterns in two languages. First phase of performance tests seems promising, and we are going to develop word and multiword alignment technique.

TRANSCRIPTION AS A TOOL FOR ANALYSIS OF PAUSES IN RUSSIAN SIGN LANGUAGE DISCOURSE.

Prozorova E.V. (zhenia-pr@yandex.ru)
Moscow State University

In this paper, we analyse pauses in Russian Sign Language discourse. In order to describe different types of pauses, we use signed discourse transcription data, which contains information on movement phases of signs and on changes in the facial expression and body posture of the signer.

INFERENCE AND ESTIMATION OF A LONG-RANGE TRIGRAM MODEL

Protasov S.V. (svp@statmt.ru)
Moscow Institute of Physics and Technology

We describe an implementation of a simple probabilistic link grammar. This probabilistic language model extends trigrams by allowing a word to be predicted not only from the two immediately preceeding words, but potentially from any preceeding pair of adjacent words that lie within the same sentence. In this way, the trigram model can skip over less informative words to make its predictions. The underlying «grammar» is nothing more than a list of pairs of words that can be linked together with. Finally, we report some experimental results using russian corpora.

NOMINALIZATIONS IN EVERYDAY SPEECH

Rozina R.I. (rarozina@yandex.ru)
Russian Language Institute, Moscow

The paper is devoted to the comparison between nominalizations in Russian everyday speech and slang on the one hand and in modern standard Russian on the other hand. Derivational bases, means of derivation, meaning, argument frame and surface behavior of nominalizations are considered. The analyses suggest that, considering the intermediate position of nominalizations between nouns and verbs, Russian colloquial and slang nominalizations are less related to motivating verbs than nominalizations in standard Russian.

ONTOLOGY EDITOR AS INTEGRATED DEVELOPMENT ENVIRONMENT

Rubashkin V. Sh. (VRubashkin@yandex.ru), Pivovarova L. M. (pivovarova@iphil.ru)
Saint-Petersburg State University

The development and usage of ontoeditor designed for operation with the knowledge model of *InTez* ontology are presented. Browsing, input, editing and other functions are discussed. The ontoeditor is compared with similar environments developed abroad.

MULTILEVEL LINGUISTIC ANNOTATION OF THE RUSSIAN SPEECH CORPUS

Ryko A.I. (aryko@mail.ru), Stepanova S.B. (stsvet_2002@mail.ru)
Laboratory of Experimental Phonetics (Institute of Philological Research of Saint-Petersburg State University), Saint-Petersburg, Russia

The paper considers multilevel linguistic annotation of the Russian Speech Corpus and its potential for description of spontaneous speech in comparison to standard language.

VARIATION IN RUSSIAN. DICTIONARY PROJECT

Savchuk S.O. (savsvetlana@mail.ru), Grishina E.A. (rudi2007@yandex.ru)
Institute of Russian Language, Moscow

The paper presents the project of the new dictionary of variants in Russian, which is supposed to be accomplished

097-Annotations:Layout 1 18.05.2008 23:04 Page 642

on the basis of the Russian National Corpus. The paper gives the preliminary description of the dictionary word list, the types of posed and solved tasks and problems.

MULTIPURPOSE DICTIONARY SUBSYSTEM FOR EXTRACTION OF SUBJECT LEXICON

Sidorova E.A. (lena@iis.nsk.su)
A.P. Ershov Institute of Informatics Systems, Russian Academy of Science

The technology intended for building of subject-oriented dictionaries and solving of various tasks of text analysis in information systems is considered. A problem of simultaneous use of several dictionaries and coordination of their contents is investigated.

PROBLEMS OF DESCRIBING COMPUTATIONAL LINGUISTICS IN ONTOLOGY OF A KNOWLEDGE PORTAL

Sokolova E. (minegot@rambler.ru), Russian State University for Humanities
Kononenko I. (irina_k@cn.ru), Institute of Informatics Systems SB RAS
Zagorulko Yu.(zagor@iis.nsk.su), Institute of Informatics Systems SB RAS

In this paper we discuss problems that emerge while developing ontology for the scientific discipline concerned with computational language, text and speech processing, that is Computational Linguistics. The problems range from defining the name and scope of the subject domain to meeting formal requirements set on the ontology specification by the knowledge portal design. Difficulties are due to the deviation of the CL from “classic” sciences like, for example, archeology, since computer for CL is not only amplification and intellectualization of modeling means. It is inherent part of the science. We consider the problems and the ontology organization.

SPEECH CORPUS OF THE RUSSIAN EVERYDAY COMMUNICATION «ONE SPEAKER’S DAY»: BASIC CONCEPTION AND CURRENT STATE

Stepanova S.B. (stsvet_2002@mail.ru), Asinovsky A.S. (a.s.asinovsky@gmail.com),
Bogdanova N.V. (nvbogdanova_2005@mail.ru), Rusakova M.V. (mvrusakova@gmail.com),
Sherstinova T.Y. (sherstinova@gmail.com)
Faculty of Philology and Arts, St. Petersburg State University, St. Petersburg, Russia

The report concerns the methodological principles elaborated for creation of the speech corpus of the Russian everyday communication “One Speaker’s Day”. The paper presents the main rules for data processing on primary stages, the description of the database, and the current state of the corpus formation.

TOWARDS HUMAN-COMPUTER INTERACTION IN NATURAL LANGUAGE

Strandson K. (krista.strandson@ut.ee), Gerassimenko O. (olga.gerassimenko@ut.ee), Kasterpalu R.(riina.kasterpalu@ut.ee), Koit M. (mare.koit@ut.ee), Rääbis A. (andriela.raabis@ut.ee)
University of Tartu, Estonia

Estonian human-human calls (directory inquiries) are analyzed with the further aim to develop a computer-human dialogue system that interacts with a user in natural language. The analysis is based on the Estonian Dialogue Corpus. Linguistic features of clients’ requests and agents’ grants are studied. A client’s initial request sets up a goal which will be achieved in collaboration with the agent. Information is given briefly by agents, using short sentences or phrases. Information-sharing sub-dialogues are initiated by both participants if either a request or a grant needs to be adjusted. A formal grammar of information dialogue is introduced in the paper. The results of the study will be implemented in two dialogue systems under development.

097-Annotations:Layout 1 18.05.2008 23:04 Page 643

RECOGNITION OF CASE SEMANTICS FOR RUSSIAN-CHINESE AUTOMATIC TRANSLATION: INSTRUMENTAL OF INSTRUMENT VS. INSTRUMENTAL OF COMPARISON

Sun Shuang (sunshuang@mail.ru), Kobozeva I. M. (kobozeva@list.ru)
Lomonosov Moscow State University

On the basis of Nirenburg & Raskin «Ontological Semantics» formal rules are proposed for recognizing semantic roles of Instrument and Similar-to (in form and in general) expressed by the instrumental case in Russian. The rules are needed for the correct translation of NP adjuncts with the head N from the class of artifacts within an AT system

THE DIRECTIONS OF INTERACTION BETWEEN HESITATION PAUSES AND KINETIC PHRASES

Sukhova N.V. (sukhova@spa.msu.ru)
Lomonosov Moscow State University

The article aims at defining a potential set of different directions in which hesitation pauses and kinetic phrases can interact. The material is a spontaneous monologue stretch of English speech. Due to a multidisciplinary approach there are seven ways detected, alongside of which the investigation of pause-kinetic interaction can be conducted.

INTEGRATION OF LINGUISTIC AND STATISTIC SEARCH METHODS IN SEARCH ENGINE “EXACTUS”

Tikhomirov I.A. (matandra@isa.ru), Smirnov I.V. (ivs@isa.ru)
Institute for Systems Analysis of RAS, Moscow

The paper considers problems of using linguistic methods of search in contemporary search engines. The features of search engine Exactus are described. The experimental evaluation of the quality of search is performed. The advantages of integration of linguistic and statistic methods are shown.

SEMANTIC FILTERS FOR THE WORD SENSE DISAMBIGUATION IN RNC: VERBS

Toldova S.Ju. (toldova@yandex.ru), Moscow State University
Kustova G.I. (galina03@mtu-net.ru), Moscow State Pedagogical University
Lyashevskaya O.N. (olesar@mail.ru), VINITI RAN

This report deals with methods of word sense disambiguation (reduction) using the information about verb argument structure. Most of the systems based on this method require specially designed resources such as WordNet, FrameNet etc. We explore the possibility to extract and use the information available from the standard dictionaries including a Verb-argument dictionary. We used a subcorpus of National corpus of Russian language that has unambiguous morphological annotation as training and testing data. The aim was to reduce the number of tags for verbs in the semantic annotation. The experiment has shown that the information extracted from dictionaries could not be used as it is. However the extracted argument structure can be used as a seed set for future training. It allows to remove rare meanings and can reduce the number of semantic tags for a verb. The further corpus training and enriching the argument structure with general semantic properties of nouns can further improve the method.

RUSSIAN CONJUNCTIONS *A TO* [LIT.: ‘AND/BUT THAT’] AND *A NE TO* [LIT.: ‘AND/BUT NOT THAT’]: WHY ARE THEY SYNONYMS IN SOME CONTEXTS?

Uryson E.V. (x-uryson@mtu-net.ru)
Institute of Russian Language, Moscow

Russian conjunctions a to [lit.: ‘and/but that’] and a ne to [lit.: ‘and/but not that’] according to their form cannot be synonyms. Yet they easily substitute for one another in some contexts. To explain this fact I analyze the element TO of these conjunctions. It derives from demonstrative/anaphoric pronoun TO(T) and in the conjunctions under discussion is not quite bleached. TO in A TO and A NE TO refers to certain fragments of a semantic structure of an utterance. The difference between the conjunctions is in the scope of TO. Compositional analysis of Russian conjunctions and particles is considered.

097-Annotations:Layout 1 18.05.2008 23:04 Page 644

DETECTING SENTENCE BOUNDARIES IN RUSSIAN

Olga Uryupina (uryupina@gmail.com)
Institute of Linguistics, Russian Academy of Science; Ashmanov and Partners

In this paper we propose a data-driven algorithm for detecting sentence boundaries in Russian. The algorithm relies on shallow features and does not require any deep syntactic knowledge. We evaluate our approach with three publicly available machine learners: C4.5, Ripper and SVM-light. The evaluation results suggest that our algorithm significantly outperforms rule-based approaches.

INTERACTION OF PRAGMATIC, AESTHETIC AND MORAL FEATURES IN THE SEMANTIC STRUCTURE OF RUSSIAN JUDGMENT ADJECTIVES

Fomchenko A.V. (degteva.anna@gmail.com), Azarova I.V. (ivazarova@gmail.com)
Saint-Petersburg State University

In the paper the core group of adjectives expressing various types of judgments are discussed. Structuring of attributive meanings in the wordnet-type thesaurus for Russian (RussNet) is described. The three facets of judgment – pragmatic, aesthetic, and moral – are considered to be fundamental.

LOCAL AND GLOBAL RULES IN SYNTAX

Zimmerling A.V. (meinmat@yahoo.com)
Moscow State University for the Humanities, MGGU/Russian State University for the Humanities, RGGU

The paper discusses word order and phrasal prosody in Russian. I claim that both phenomena can be described in terms of two successive sets of rules — local rules vs. global rules. Combinations of these two sets of rules are typical of multilayer language models and for algorithmic generation of complex structural objects in formal grammars. Modern Russian applies to a highly formalized rule of choosing the locus of the main phrasal accent: the hierarchy of potential accent bearers is a mirror image of the grammatical hierarchy of arguments and adjuncts. The order of communicative constituents in Russian is governed by 7-8 Linear-Accent Transformations (LA-transformations). LA-transformations are Movement rules, which both operate on constituent order and change accent markings of communicative constituents. In the preceding Russian linguistic tradition (cf. Paducheva and Yanko) LA-transformations are defined as Context-Sensitive rules, which makes word order calculus impossible. I discuss the possibility to reformulate LA-transformations as pairs of the type <Active & Remnant Movement> and offer an analysis compatible with Mildly Context-Sensitive Grammars, e.g. Stablerian Minimalist Grammars.

ALGORITHM OF THE INTONATION MARKING OF NARRATIVE SENTECES FOR TTS SYNTHESIS

Tsirulnik L.I. (L.Tsirulnik@newman.bas-net.by), Lobanov B.M. (Lobanov@newman.bas-net.by),
Sizonov O.G. (Osizonov@yahoo.co.uk)
United Institute of Informatics Problems on National Academy of Science of the Republic of Belarus

The paper presents an algorithm of segmentation into phrases and intonation tagging of narrative sentences. The algorithm takes into account the positional and combinatory prosodic factors. The use of the proposed algorithm in TTS synthesis system provides an elimination of so called “second degree of monotony” in synthesized speech.

BORDERLINES BETWEEN EMOTIONAL INTERJECTIONS AND MODAL PARTICLES

Sharonov I.A. (igor_sharonov@mail.ru)
Russian State University for Humanities

The research is aimed to distinguish interjections and participles with a help of syntactic, semantic and pragmatic criteria. The word should be regarded as interjection, if it is syntactically autonomous, spontaneous and not addressed reaction to linguistic, and also to extra-linguistic stimulus

097-Annotations:Layout 1 18.05.2008 23:04 Page 645

VERBS OF GOING DOWN: SEMANTICS AND COMPATIBILITY

Shemanaeva O.Yu. (shemanaeva@yandex.ru)
Russian State University for Humanities

Russian verbs of going down are described in this paper. The relevant parameters of adequate semantic description are shown, for example the control of the subject, the speed of movement, the layer in which the subject is being put. Three main metaphorical extensions – BAD IS DOWN, the large amount of something and the disappearance from sight are being discussed.

"WE" AND "OTHERS": THE SIMULATION OF UKRAINIAN SPEECH IN RUSSIAN JOKES

Shmeleva E.Y. (eshkind@mail.ru), Shmelev A.D. (Smelev.Alexei@gmail.com)
Institute of Russian Language, Moscow

Simulating Ukrainian speech, making fun of funny-sounding Ukrainian words and names are unmistakable sings that jokes about Ukrainians are produced in the Russian linguistic environment. The paper aims at revealing links between typical joke plots, "linguistic masks" of the characters, and ethnic stereotypes.

VOXFORGE.ORG FREE SPEECH CORPUS

Shmyrev N.V. (nshmyrev@yandex.ru)
SRISA RAS, Moscow

We discuss the work on building the first free speech database for recognition systems. This report reviews free speech sources, processing technique and problems related to the collection of the big multilingual speech database.

SET OF RECOGNIZABLE WORDS AS COMPRESSION TEXTS (WITH COMPARISON OF KEY-WORD SET)

Iagounova E.V. (iagounova_elena@mail.ru)
St.Petersburg State University

Main characteristics of set of recognizable words (in perception text in white noise) have been described in terms of compression texts (with comparison of key-word set). Results of reconstruction text with the set words are analyzed with reference of discovering main characteristics of the set. One of the most finding is the dependence of sense structure of a text on following text parameters: professional vs. fiction and dynamic vs. static.

PROSODY IN A DICTIONARY, AND A DICTIONARY OF PROSODIC IDIOMS

Yanko T.E. (tanya_yanko@list.ru)
Institute of Linguistics (Russian Academy of sciences)

Representing prosodic data in a dictionary raises two problems: to account for limitations on the communicative and prosodic application of words and constructions by their definitions or functions in discourse; to collect idiomatic illocutions and their prosodic parameters in a prosodic dictionary.

CHOOSING LANGUAGE IN INTERNET CONVERSATION BETWEEN RUSSIANS AND ESTONIANS

Oja Anni (anni.oja@tlu.ee)
Tallinn University

Current study examines interlingual communication in Estonian web-portal rate.ee. First conversations between Estonians and Russians are viewed in order to see the factors in choosing language for first conversation act (conversations are normally strings of picture comments). Most of these factors are related to situation (who are the participants, how it is more comfortable to communicate, what is the purpose), but some things are learned unintentionally via community of practice, generally environment-related unwritten rules of politeness and polite language choices with equipment of suitable vocabulary.

097-Annotations:Layout 1 18.05.2008 23:04 Page 646

SYMMETRY AND SYMMETRICAL PREDICATES

Partee B.H. (partee@linguist.umass.edu)
University of Massachusetts, Amherst, MA, USA and RGGU, Moscow

A goal of this paper is to analyze the differences between mathematical definitions of symmetry and a concept of symmetry that would fit best with observed linguistic generalizations. This requires a closer look at some aspects of the linguistic behavior of symmetric and non-symmetric predicates.

ARTIFICIAL COMPANIONS AS A NEW KIND OF DIALOGUE INTERFACE TO THE FUTURE INTERNET

Yorick Wilks (yorickwilks@googlemail.com)
University of Sheffield, UK

This paper seeks to connect the future of the Internet to a new, even though relatively underdeveloped, technology, that of computer speech and language and its embodiment, in a concept I shall call an Artificial Companion. Before moving to describe the integration that constitutes the Companion, we must first mention two technologies, not only in their own right but because, in each case, there have been misunderstandings about their achievements and goals. They are:

Language and speech technology

Agents and the Semantic Web

The first of these is Berners-Lee's [Berners-Lee et al., 2001] vision of how the Internet will change, and it is to that new Internet we intend the Companion as the human interface, on the ground that without it the Internet may get harder and not easier to use, and we shall return to the Semantic Web at the end of this paper. The second notion above is that agents will change from transitory software entities that e.g. locate a cheap camera on the internet, to more permanent social Companion entities that deal with a user through dialogue over a long period, learn his or her needs and preferences and elicit large quantities of life data though conversation.

BERMUDAN TRIANGLE INTERACTION – COMMUNICATION – CONTACT

Narin'yani A.S. (narin@aha.ru)
«IntelliTek» Company

The triangle denoted in the title is representing a poorly defined part of the Knowledge System. On the one hand, an enormous amount of works are dedicated to this territory. On the other hand it remains to be too large and ill investigated for viewing it as a whole and defining some sufficiently concrete boundaries between denoting it conceptions. To a considerable degree this is related to the fact that those notions are basic ones and consequently are not defined sufficiently clearly. At the same time they are closely connected with the notion of information, which has been defined more than once in different ways, so it's not too productive in the context intended. In the report an attempt is undertaken to outline at least a rough map of that terminological triangle.

098-Autors:Layout 1 18.05.2008 23:04 Page 647

Авторский указатель

098-Autors:Layout 1 18.05.2008 23:04 Page 648

099-Vyhod:Layout 1 18.05.2008 23:04 Page 649

Научное издание

Компьютерная лингвистика и интеллектуальные технологии
по материалам ежегодной Международной конференции «Диалог» (2008)
Периодическое издание, выпуск 7 (14)

Утверждено к печати Ученым советом Института проблем информатики РАН

Компьютерная верстка А.Н. Каллистов

Отпечатано с готового оригинал-макета в издательском центре

Российского государственного гуманитарного университета

125993 Москва, Миусская пл., 6

Подписано к печати 19.05.08 г. Формат 60x84/8.
Усл. печ. л. 87,1. Уч.-изд. л. 112.
Бумага офсетная. Тираж 200 экз. Заказ № 104.