Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной Международной конференции «Диалог» (2012)

Выпуск 11

В двух томах

Том 1. Основная программа конференции

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2012)

Issue 11

Volume 1 of 2. Main conference program

УДК 80/81; 004
ББК 81.1
К63

Программный комитет конференции выражает
искреннюю благодарность Российскому фонду фундаментальных
исследований за финансовую поддержку,
грант № 12-06-06045-г

Редакционная коллегия: *А. Е. Кибрик* (главный редактор), *В. И. Беликов, И. М. Богуславский, Б. В. Добров, Д. О. Добровольский, Л. М. Захаров, Л. Л. Иомдин, И. М. Кобозева, Е. Б. Козеренко, М. А. Кронгауз, Н. И. Лауфер, Н. В. Лукашевич, Д. Маккарти, Й. Нивре, Г. С. Осипов, В. Раскин, И. В. Сегалович, В. П. Селегей, Э. Хови, С. А. Шаров*

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая–3 июня 2012 г.). Вып. 11 (18): В 2 т. Т. 1: Основная программа конференции. — М.: Изд-во РГГУ, 2012.

Сборник включает 78 докладов международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2012», 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 теоретической и прикладной лингвистики и интеллектуальных технологий.

© Редколлегия сборника «Компьютерная лингвистика и интеллектуальные технологии» (составитель), 2012
© Российский государственный гуманитарный университет, 2012

Предисловие

11-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит материалы 18-й Международной конференции «Диалог». Для сборника было отобрано 78 докладов, охватывающих различные направления исследований в области компьютерного моделирования и анализа естественного языка. В настоящем сборнике представлены:

- Лингвистическая семантика и семантический анализ
- Формальные модели языка и их применение
- Теоретическая и компьютерная лексикография
- Методы оценки (evaluation) систем анализа текстов
- Корпусная лингвистика; создание, применение, оценка корпусов
- Интернет как лингвистический ресурс; лингвистические технологии в Интернете
- Извлечение знаний из текстов
- Компьютерный анализ документов: реферирование, классификация, поиск
- Автоматический анализ тональности текстов
- Модели синтаксиса для задачи синтаксического парсинга
- Машинный перевод
- Модели общения, коммуникация, диалог и речевой акт
- Анализ и синтез речи

«Диалог» – самая крупная конференция по компьютерной лингвистике, проводимая в России. Принципиальной особенностью конференции является пристальное внимание к технологиям автоматического анализа языка, основанным на лингвистических моделях. Именно этим объясняется и состав участников, и программа конференции, в которой соседствуют теоретические и прикладные исследования. На «Диалоге» представлены и работы, сделанные в рамках статистических подходов, и гибридные системы, что позволяет, в частности, сравнивать полученные результаты.

В этом году на «Диалоге» подводятся итоги двух тестовых испытаний: систем анализа тональности текстов (специальная дорожка РОМИПа) и систем синтаксического анализа (2-й этап Форума по тестированию систем автоматического анализа текстов). Выбор именно таких направлений тестирования оказался очень полезным, поскольку предложенные участникам задачи в этих двух испытаниях оказались принципиально разными с точки зрения методов оценки.

Тестирование анализа тональности представляет собой образцовую прикладную задачу компьютерной лингвистики с простыми и эффективными методами оценки. При анализе русскоязычных тестовых коллекций разработчики имели возможность использовать любые существующие или новые методы – и опубликованные результаты тестирования позволяют оценить их сравнительный потенциал.

Тестирование систем синтаксического анализа было направлено не на анализ какой-то частной прикладной задачи, а на качество полученной синтаксической разметки как таковое. Для русского языка подобное сравнительное тестирование проводилось впервые, и задача унификации используемых разработчиками синтаксических моделей оказалась весьма сложной.

Опубликованные в сборнике результаты показывают, с какими проблемами пришлось столкнуться как разработчикам, так и организаторам тестирования. Тем не менее, этот новый шаг в разработке методов содержательного тестирования систем автоматического анализа русского языка представляется весьма поучительным и очень важным для «Диалога» с точки зрения его глобальной задачи.

В сборник включены итоговые статьи организаторов двух испытаний и большая часть комментирующих статей участников. Полностью статьи участников публикуются на сайте конференции и в её электронных материалах.

Постоянной доминантой «Диалога» является корпусная тематика. В прошлом году началась интенсивная дискуссия о методах оценки адекватности корпусов и применяемых методик корпусных исследований для решения различных задач лингвистики и лексикографии. Эта дискуссия имела продолжение после конференции и получила новое развитие в работах, представленных в этом году. В сборнике публикуются статьи, посвященные анализу практики применения существующих корпусов и проектам разработки корпусов нового типа.

Традиционно важное место в программе «Диалога», выросшего из междисциплинарных семинаров «Модели общения», занимают исследования звучащей речи и коммуникативных стратегий, в том числе невербальных.

Несмотря на широту тематики докладов этого года, они тем не менее не отражают полной картины направлений конференции. Более полное представление о проблематике «Диалога» можно получить на сайте конференции www.dialog-21.ru, где опубликованы обширные электронные архивы «Диалогов» прошлых лет.

Программный комитет конференции «Диалог»
Редакционная коллегия ежегодника «Компьютерная лингвистика и интеллектуальные технологии»

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
- Филологический факультет МГУ

Конференция проводится при поддержке Российской ассоциации искусственного интеллекта.

Международный программный комитет

Буате Кристиан	Гренобльский университет
Богуславский Игорь Михайлович	Политехнический Университет Мадрида
Гельбух Александр Феликсович	Национальный политехнический институт, Мехико
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН
Кибрик Александр Евгеньевич	Филологический факультет МГУ
Кобозева Ирина Михайловна	Филологический факультет МГУ
Козеренко Елена Борисовна	Институт проблем информатики РАН
Корбетт Гревил	University of Surrey, UK
Кронгауз Максим Анисимович	Институт лингвистики РГГУ
Лукашевич Наталья Валентиновна	НИВЦ МГУ
Маккарти Диана	Lexical Computing Ltd., UK
Мельчук Игорь Александрович	Монреальский университет
Нивре Йоаким	Уппсальский университет
Ниренбург Сергей	Университет Нью-Мексико
Осипов Геннадий Семёнович	Институт программных систем РАН
Попов Эдуард Викторович	РосНИИ информационной техники и САПР
Раскин Виктор	Purdue University, USA
Сегалович Илья Валентинович	Компания Яndex
Селегей Владимир Павлович	Компания ABBYY
Хови Эдуард	University of Southern California
Шаров Сергей	University of Leeds, UK

Организационный комитет и Редсовет

Селегей Владимир Павлович, *председатель*	Компания ABBYY
Беликов Владимир Иванович	Институт Русского Языка им. В. В. Виноградова
Добров Борис Викторович	НИВЦ МГУ
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН
Кобозева Ирина Михайловна	Филологический факультет МГУ
Козеренко Елена Борисовна, *руководитель проекта по гранту РФФИ 12-06-06045-г*	Институт проблем информатики РАН
Лауфер Наталия Исаевна	ООО «проФан Продакшн»
Ляшевская Ольга Николаевна	Universitetet i Tromsø, Norway
Соколова Елена Григорьевна	Институт лингвистики, РГГУ
Толдова Светлана Юрьевна	Филологический факультет МГУ
Шаров Сергей	University of Leeds, UK

Секретариат

Талис Валентина Львовна, *секретарь оргкомитета, редактор сайта*	Компания ABBYY
Мытникова Татьяна Александровна, *координатор*	Компания ABBYY
Морозова Юлия Игоревна	ИПИ РАН

Рецензенты

Августинова Таня
Азарова Ирина Владимировна
Апресян Валентина Юрьевна
Баранов Анатолий Николаевич
Беликов Владимир Иванович
Богданова Наталья Викторовна
Богданов Алексей Владимирович
Бонч-Осмоловская Анастасия
Браславский Павел Исаакович
Гельбух Александр Феликсович
Горностай Татьяна Александровна
Губин Максим Вадимович
Даниэль Михаил Александрович
Добров Борис Викторович
Добровольский Дмитрий Олегович
Добрынин Владимир Юрьевич
Зарецкая Елена Наумовна
Захаров Леонид Михайлович
Зуев Константин Алексеевич
Иомдин Борис Леонидович
Иомдин Леонид Лейбович
Кибрик Андрей Александрович
Кобозева Ирина Михайловна
Козеренко Елена Борисовна
Крейдлин Григорий Ефимович
Кронгауз Максим Анисимович
Лапшин Владимир Анатольевич
Лахути Делир Гасемович
Левонтина Ирина Борисовна
Лобанов Борис Мефодьевич
Лукашевич Наталья Валентиновна
Ляшевская Ольга Николаевна
Маккарти Диана
Новицкий Валерий Игоревич
Нивре Йоахим
Пазельская Анна Германовна
Плунгян Владимир Александрович
Раскин Виктор
Рахилина Екатерина Владимировна
Савельев Василий Евгеньевич
Селегей Владимир Павлович
Сокирко Алексей Викторович
Соколова Елена Григорьевна
Тестелец Яков Георгиевич
Тихомиров Илья Александрович
Урысон Елена Владимировна
Филиппова Екатерина Александровна
Хорошевский Владимир Федорович
Циммерлинг Антон Владимирович
Хови Эдуард
Шаров Сергей Александрович
Юдина Мария Владимировна
Янко Татьяна Евгеньевна

Содержание*

Раздел I.
Основная программа конференции

Апресян В. Ю.
***Тут* как показатель темпоральной близости** 1

Arkhipov A. V., Zakharov L. M., Krivnova O. F., Kodzasov S. V., Lebedev A. A.
Russian intonation corpus: a preliminary report 18

Баранов А. Н., Вознесенская М. М., Добровольский Д. О.,Киселева К. Л., Козеренко А. Д.
Проект частотного словаря русских идиом 28

Беликов В. И., Селегей В. П., Шаров С. А.
Пролегомены к проекту Генерального интернет-корпуса русского языка (ГИКРЯ) 37

Бочаров В. В., Алексеева С. В., Грановский Д. В., Остапук Н. А., Степанова М. Е., Суриков А. В.
Сегментация текста в проекте «Открытый корпус» 51

Богданов А. В.
Описание гэппинга в системе автоматического перевода 61

Богданова Н. В.
О проекте словаря дискурсивных единиц русской речи (на корпусном материале) 71

Большаков И. А., Большакова Е. И.
Автоматический морфоклассификатор русских именных групп 81

Борисова Е. Г.
Роль дискурсивных слов в управлении пониманием текста 93

Chistikov P. G., Korolkov E. A.
Data-driven speech parameter generation for Russian text-to-speech system 103

Даниэль М. А., Зеленков Ю. Г.
НКРЯ как инструмент социолингвистических исследований. Эпизод IV. Пол говорящего и длина реплики 112

Давыдов А. Г., Киселёв В. В., Кочетков Д. С., Ткаченя А. В.
Выбор оптимального набора информативных признаков для классификации эмоционального состояния диктора по голосу 122

---

* Доклады внутри каждого раздела упорядочены по фамилии первого автора в соответствии с порядком английского алфавита.

Деликишкина Е. А., Федорова О. В.
Влияние фактора синтаксической роли антецедента на разрешение референциальной неоднозначности в русском языке 129

Добровольский Д. О., Левонтина И. Б.
О синонимии фокусирующих частиц (на материале немецкого и русского языков) 138

Добрушина Н. Р.
Сентенциальные дополнения с союзом *чтобы* 150

Герасименко О. А.
Функции частиц обратной связи в телефонном диалоге (на примере лексем *а*, *ага* и *гм*) 162

Гришина Е. А.
Автодейксис: основные типы и значения 173

Гурин Г. Б., Беликова А. Е.
Методика оценки конвенциональности метафорических выражений: от интуитивистских критериев к операциональным 187

Hetsevich Yu. S., Hetsevich S. A., Lobanov B. M.
Belarusian and Russian linguistic processing modules for the system NooJ as applied to text-to-speech synthesis 198

Иомдин Б. Л., Лопухина А. А., Пиперски А. Ч., Киселева М. Ф., Носырев Г. В., Рикитянский А. М., Васильев П. К., Кадыкова А. Г., Матиссен-Рожкова В. И.
Словарь бытовой терминологии: новые проблемы и новые методы 213

Kashkin E. V., Reznikova T. I., Pavlova E. K., Luchina E. S.
Verbs describing sounds of inanimate objects: towards a typology 227

Кибрик А. А., Линник А. С., Добров Г. Б., Худякова М. В.
Оптимизация модели референциального выбора, основанной на машинном обучении 237

Кюсева М. В., Рыжова Д. А., Холкина Л. С.
Прилагательные *тяжелый* и *легкий* в типологической перспективе 247

Клыгина Е. А., Крейдлин Г. Е.
База данных «Тело и телесность в языке и культуре (идеология, структура и наполнение)» 256

Klyueva N. M.
Some differences between Czech and Russian: a parallel corpus study 268

Kobozeva I. M., Lukashevich N. Iu.
Human characters through the prism of adverbs 277

Kostyrkin A. V., Panina A. S., Reznikova T. I., Bonch-Osmolovskaia A. A.
Constructing a lexico-typological database (for a study of pain predicates) 288

Kotov A., Budyanskaya E.
The Russian emotional corpus: communication in natural emotional situations 296

Краснова Е. В., Смирнова Н. С.
Региональные предпочтения в произношении некоторых русских и заимствованных слов 307

Кравченко А. Н.
Генерация шаблонов оценочных выражений на основе неразмеченного текста 319

Krylov S. A.
The general corpus of the modern Mongolian language and its structural-probabilistic model 331

Крылова Т. В.
Слова со значением «лекарственное вещество»: вчера и сегодня 342

Кустова Г. И.
Семантические модели и семантические функции отпричастных прилагательных 352

Kutuzov A. B., Kunilovskaya M. A., Oschepkov A. Y., Chepurkova A. Y.
Russian learner parallel corpus as a tool for translation studies 362

Ляшевская О. Н., Митрофанова О. А., Грачкова М. А., Романов С. В., Шиморина А. С., Шурыгина А. С.
К построению инвентаря русских именных конструкций 370

Людовик Т. В.
Влияние речевых характеристик дикторов на точность автоматического распознавания речи 383

Liusina V. S.
Oral speech and instant messaging: the impact of the communication channel on dialogue structure 393

Малкова А. С.
О структуре тезауруса пословиц 404

Маничева Е. С., Дрейзис Ю. А., Селегей В. П.
Разработка лексико-семантического словаря китайского языка для многоязычной системы анализа текстов 418

Михеев М. Ю.
Шолохов или все-таки Крюков? Неформальные процедуры при установлении авторства «Тихого дона» 431

Mihkla M., Hein I., Kalvik M., Kiissel I., Tamuri K., Sirts R.
Estonian speech synthesis: applications and challenges 443

Moldovan D.
Representing and reasoning for explicit, implicit and implicated textual information 454

Mukhin M., Braslavski P.
What do people ask the community question answering services and how do they do it in Russian? 464

Нехай И. В.
Применение n-грамм и других статистик уровня символов и слов для семантической классификации незнакомых собственных имен 477

Nokel M. A., Bolshakova E. I., Loukachevitch N. V.
Combining multiple features for single-word term extraction 490

Орехов Б. В., Галлямов А. А.
Башкирский интернет: лексика и прагматика в количественном аспекте 502

Осьмак Н. А.
О способах лексикографического описания русской спонтанной речи (корпусное исследование) 510

Падучева Е. В.
К семантике коммуникативной структуры: исходные структуры и линейно-акцентные преобразования 522

Поляков А. Е.
Проблемы и методы анализа русских текстов в дореформенной орфографии 536

Савчук С. О.
Варианты родовой принадлежности в группе имен существительных pluralia tantum в русском языке 548

Schumann A.-K.
Towards the automated enrichment of multilingual terminology databases with knowledge-rich contexts 559

Семенова С. Ю.
Русское имя параметра: метафорические и метонимические процессы 568

Шарапов Р. В., Шарапова Е. В.
Исследование плагиата в работах студентов 578

Шмелев А. Д.
Корпус или эксперимент? 587

Sokolova E. G., Kononenko I. S.
Russian-English thesaurus on computational linguistics 598

Solomennik A. I., Chistikov P. G.
Automatic generation of text corpora for creating voice databases in a Russian text-to-speech system 607

Solov'ev A. N., Antonova A. Ju., Pazel'skaia A. G.
Using Sentiment-analysis for text information extraction 616

Урысон Е. В.
Союзы, коннекторы и теория валентностей 627

Валиахметова А. Р.
Управление иноязычных существительных — названий лиц по профессии в современном Русском языке 638

Ягунова Е. В.
Экспериментальный подход к определению коммуникативной структуры текста 652

Yanko T. E.
Extracting subjective information from spoken discourse: the case for prosodic emphasis 664

Загорулько М. Ю., Кононенко И. С., Сидорова Е. А.
Система семантической разметки корпуса текстов в ограниченной предметной области 674

Зализняк А. А., Шмелев А. Д.
Русское *разочарование* в европейском контексте: история и современность 684

Занегина Н. Н.
Словосочетания-записанные-через-дефис как способ категоризации и новый прием языковой выразительности в русском языке 696

Zangenfeind R.
Towards a system of syntactic dependencies of German 706

Zhila A., Gelbukh A.
Exploring context clustering for term translation 716

Zimmerling A.
A unified analysis of clitic clusters in world's languages 726

Abstracts 739

Авторский указатель 760

Раздел I.
Основная программа конференции

ТУТ КАК ПОКАЗАТЕЛЬ ТЕМПОРАЛЬНОЙ БЛИЗОСТИ[1]

Апресян В. Ю. (valentina.apresjan@gmail.com)

Высшая школа экономики, Москва, Россия
Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Метафора ‘пространство-время’ — одна из универсалий семантической типологии (особенно на материале предлогов и наречий), однако в дейксисе этот перенос происходит значительно реже (так, здесь не развивает значение ‘сейчас’, там не развивает значение ‘тогда’ и т. п.). Возможно, в этом нет необходимости, т. к. у временного дейксиса существуют свои собственные независимые лексические средства. Однако и в дейксисе эта метафора не столь экзотична, как может показаться на первый взгляд, что можно продемонстрировать на примере полисемии пространственного дейктического наречия тут, которое развивает значение близости к моменту речи. Однако значение временного тут отличается от значений классических маркеров момента речи в русском языке — наречий сейчас и теперь. Значение тут, которое можно определить как значение актуальности, сочетает в себе следующие компоненты: (а) период времени, включающий в себя момент речи, а также такие моменты до и после него, которые достаточно к нему близки, чтобы сохранять с ним связь; (б) ситуацию, разворачивающуюся в том же физическом или ментальном пространстве, в котором находится говорящий, и участником или наблюдателем которой он является. Помимо семантических особенностей, тут в значении актуальности отличается от сейчас грамматически, коммуникативно и просодически: тут имеет полуграмматикализованный статус, является клитикой, не может входить в рему или контрастную тему, не может нести на себе ударения.

Ключевые слова: пространственный и временной дейксис, момент речи, семантика, просодия, ситуация, коммуникативная функция, грамматикализация

---

[1] Данная работа выполнена при поддержке следующих грантов: гранта Программы фундаментальных исследований ОИФН РАН «Язык и литература в контексте культурной динамики», гранта РГНФ № 10-0400273а «Подготовка первого выпуска Активного словаря русского языка», а также гранта НШ-6577.2012.6 для поддержки научных исследований, проводимых ведущими научными школами РФ.

1. Введение

Хорошо известно, что во многих языках идея времени передается при помощи слов со значением пространства.

Время-пространство — одна из самых универсальных и хорошо изученных языковых метафор, которая проявляется, среди прочего, в развитии у пространственных предлогов временных значений. Однако в области дейксиса время и пространство существуют более обособленно, а именно, время выражается собственными дейктическими средствами; ср. наречия *сейчас, теперь, тогда*, которые не являются производными от пространственных дейктических слов со значением близости или удаленности, соответственно. Разграниченность пространственного и временного дейксиса достаточно универсальна — использование пространственных дейктических слов для обозначения времени как минимум не является распространенной тенденцией в европейских языках (*здесь* не значит *сейчас*, *here* не значит *now*, *ici* не значит *maintenant*, *hier* не значит *jetzt* и т. п.). Ситуации, когда временной дейксис выражается через пространственный, как это возможно, например, в норвежском языке[2], представляются скорее экзотическими.

В русском языке основными дейктическими словами со значением темпоральной близости к моменту речи являются *сейчас* и *теперь*. *Сейчас*, более универсальное слово, может употребляться и для референции к моменту речи, и к более абстрактному настоящему, и для выражения недавнего прошлого, и для выражения близкого будущего:

(1а) *Мы сейчас читаем «Преступление и наказание»* [в момент речи, например, в данный момент на уроке]

(1б) *Мы сейчас читаем «Преступление и наказание»* [в период времени, в котором говорящий себя мыслит, например, в школе в течение последних двух недель]

(1в) *Мы сейчас читали «Преступление и наказание» и обсуждали отношение Достоевского к либералам* [незадолго до момента речи]

(1г) *Мы прочли «Записки из подполья» и сейчас будем читать «Преступление и наказание»* [вскоре после момента речи].

---

[2] Ср. [Fretheim 2010] о показателе пространственной близости *her* в значении показателя недавнего прошлого]: *Jeg var i London* ***her*** *i oktober* ‘Я был в Лондоне **здесь** в октябре’ [в октябре, ближайшем к моменту речи] vs. *Jeg var i London i oktober* ‘Я был в Лондоне в октябре’ [в октябре любого года].

Дейктическое наречие *теперь* обладает спецификой относительно семантически более универсального *сейчас*, а именно, указывает на то, что нынешней ситуации предшествовала какая-то другая[3]; ср. *сейчас* и *теперь* в следующих фразах

(2а) *Мне сейчас неудобно с ним разговаривать* [в данный момент]

(2б) *Мне теперь неудобно с ним разговаривать* [после того, что произошло до этого]

Таким образом, русский язык обладает разными средствами выражения идеи настоящего, а также идей недавнего прошлого и близкого будущего и, казалось бы, не имеет нужды в дополнительных подобных средствах. Однако оказывается, что в русском, как и в норвежском, возможно использование пространственного дейксиса со значением близости для обозначения близости временной. Набором временных значений обладает дейктическое слово *тут*, причем, хотя словари русского языка отмечают некоторые временные значения, ни в одном словаре не описано значение, в котором *тут* выражает близость к моменту речи, т. е. является синонимом *сейчас*.

2. *Тут* как показатель близости к моменту речи

В качестве показателя близости к моменту речи *тут* имеет различные употребления, весьма сходные с употреблениями *сейчас*, причем часто возможна взаимозамена этих двух показателей:

(3а) *Я тут <сейчас> на совещании, я тебе попозже позвоню* [в момент речи]

(3б) *Мы тут <сейчас> открываем новый филиал в Новосибирске* [в период времени, в котором говорящий себя мыслит]

(3в) *Я тут <сейчас> с ним говорила по телефону* [незадолго до момента речи]

(3г) *Я тут <сейчас> буду заполнять налоговую декларацию* [вскоре после момента речи]

Однако, как будет показано ниже, *тут* не является полным синонимом *сейчас*, имеет собственную (и весьма существенную) семантическую, грамматическую и просодическую специфику, и выражает такие значения, которые не «обслуживаются» никакими другими языковыми средствами выражения темпоральности в русском языке.

В следующих двух секциях рассматриваются прочие темпоральные значения *тут* и демонстрируются их отличия от *тут* в значении близости к моменту речи, а также рассматриваются сходства и различия между *тут* и *сейчас*.

---

[3] См. об этом [Мельчук 1995], [Урысон 2004].

2.1. Отличия *тут* как показателя близости к моменту речи от других временных значений *тут*

МАС выделяет у наречия *тут* следующие не-локативные значения (приводятся только те из них, которые имеют потенциальное отношение ко времени):

1) В это время, тогда. *[Девочка] вместе с братцем прибежала домой. А тут и отец с матерью пришли.* А. Н. Толстой, Гуси-лебеди.

2) || В этом случае, при этих обстоятельствах. — *Он советует ехать на воды за границу. Что ты скажешь? — Что же мне сказать? Тут, я думаю, голос доктора важнее моего.* И. Гончаров, Обыкновенная история.

4) *в знач. частицы.* Употребляется для приступа к речи, при переходе к какой-л. теме, примеру. — *Тут к нам ездит один ферт со скрипкой и пиликает.* Чехов, Черный монах. *[Яковлев:] Тут сейчас человек должен прийти.* М. Горький, Фальшивая монета. *Тут как-то, сидя на койке у себя со свечкой, сосчитал Брыкин сумму богатства своего.* Леонов, Барсуки.

Из этих трех значений первое выражает собственно идею времени, второе указывает на соотношение некоторых обстоятельств и ситуаций, а именно на то, что при наличии обстоятельств или ситуации A1 имеет место ситуация A2, т. е. по семантике оно является скорее условным. Временной компонент здесь присутствует очень неявно. Он проявляется, во-первых, в том, что всякая ситуация и всякие обстоятельства имеют привязку к какому-то времени и месту (в этом смысле это значение локативнотемпоральное), а во-вторых, в том, что наличию некоторой ситуации **предшествует** наличие некоторой другой ситуации и обстоятельств. При этом сами обстоятельства обычно не эксплицируются в пределах высказывания с *тут*, о них становится известно из контекста:

(4а) — *Что делать, если посаженные черенки завяли <если вашу машину поцарапали на парковке*; *если вы нашли на себе клеща>? — Тут надо действовать без промедления*

(4б) — *Что делать, если вы хотите продать квартиру <если вы попали в аварию*; *если вы хотите зарегистрировать новую фирму>? — Тут без помощи юриста не обойтись*

В дальнейшей части статьи это значение разбираться не будет. Что касается четвертого МАСовского значения, то оно, как представляется, соответствует выделяемому в данной статье *тут* в значении близости к моменту речи, хотя в МАСе толкуется как не-временное. В данной статье рассматриваются МАСовские значения 1) и 4), оба из которых содержат указание на близость во времени. Однако, как будет показано ниже, эти два временных значения различны и в каком-то смысле даже семантически противоположны. Рассмотрим сначала

МАСовское значение ‘в это время, тогда’. В конце следующего подраздела предлагается полное толкование этого значения, однако пока что уточним его синоптическое определение: ‘в момент или после момента А1’.

2.1.1. *Тут* в значении ‘в момент или после момента А1’

I. Во-первых, в этом значении *тут* выражает таксис, а именно указывает на временную соотнесенность двух ситуаций. *Тут* вводит ситуацию, которая возникает п о с л е наступления какой-то другой ситуации:

(5а) *Я уже собиралась уходить* [А1]*, и тут мне наконец позвонили* [А2]

(5б) *Иду я вчера по улице* [А1]*, и тут ко мне подходит какой-то незнакомый человек* [А2]

(5в) *Все зрители расселись по местам* А1]*, в зале потушили свет* А1]*, и тут раздался чей-то страшный крик* [А2]

II. Во-вторых, в отличие от *потом*, *тогда* и даже таких наречий как *вскоре* и *сразу же*, которые допускают наличие временного промежутка между А1 и А2, наречие *тут* указывает на непосредственное, н е м е д л е н н о е следование ситуаций (5а), (5в) или же на «наложение» ситуаций друг на друга (5б), когда вторая начинается в то время, когда еще продолжается первая. Две ситуации должны субъективно мыслиться как часть одного, небольшого временного пространства, или же одного короткого периода времени. В этой особенности временного *тут* проявляется семантическая специфика исходного, локативного значения близости. Ср. естественность (6а) и странность (6б):

(6а) *Они наняли хорошего адвоката и вскоре <сразу же> — не больше, чем через неделю — все проблемы были улажены*

(6б) ?*Они наняли хорошего адвоката и тут — не больше, чем через неделю — все проблемы были улажены*

III. В-третьих, *тут* указывает не только на единство времени, но и на единство м е с т а, т.е. А1 и А2 не могут происходить в разных пространствах (в этом смысле *тут* противопоставлено таким единицам как *после того как*, где А1 и А2 представляются более независимыми ситуациями). При этом пространства могут быть физически разными, однако они обязательно должны быть объединены в сознании говорящего в некое единое наблюдаемое мысленное пространство. Ср. следующие примеры: естественный (7а) со словом *тут*, где обе ситуации происходят в одном физическом пространстве; естественный (8) со словом *тут*, где разные физические пространства составляют единое ментальное целое; естественный (7б) без слова *тут*, где речь идет о физически и ментально различающихся пространствах, и, наконец, прагматически

неестественный пример (7в) со словом *тут*, где физически различные пространства когнитивно не объединены:

(7а) *Ураган сорвал крышу с сарая и тут прекратился*

(7б) *Ураган прошел через штат Иллинойс и прекратился в Новой Англии*

(7в) ??*Ураган прошел через штат Иллинойс, и тут прекратился в Новой Англии*

(8) *В 1830 Алжир был захвачен французами, и тут во Франции разразилась революция*

Таким образом, в *тут* заложено представление о наблюдателе, зрителе (он может быть также участником событий), который наблюдает некую последовательность событий, сценическое действие, условием которого является единство времени и места. Неудивительны поэтому видо-временные характеристики глаголов, которые употребляются с *тут*: в контексте *тут* глаголы часто используется в настоящем сценическом, для создания эффекта присутствия и репортажа с места событий: *И тут он входит*.

В отношении глагольных форм, с которыми сочетается *тут*, интересен контраст между активным и пассивным залогом; *тут* плохо сочетается с пассивом в форме СОВ ПРОШ, хотя сочетается с активным залогом в СОВ ПРОШ, а также с пассивным залогом в форме настоящего исторического или сценического. На синонимы *тут*, например, на наречие *вскоре*, эти ограничения не распространяются. Ср.

(9а) *Мы купили машину, и тут рядом с нами построили новую станцию метро*

(9б) *Мы покупаем машину, и тут рядом с нами строят новую станцию метро*

(9в) ?*Мы купили машину, и тут рядом с нами была построена новая станция метро*

(9г) *Мы купили машину, а вскоре рядом с нами была построена новая станция метро*

По-видимому, в значение пассива СОВ ПРОШ входит некая объективизация и отстраненность, которая мешает мыслить событие, описываемое им, как происходящее актуально и разворачивающееся перед реальным или мысленным взором наблюдателя.

IV. Наконец, *тут* в этом значении вводит элемент некоторой неожиданности второй ситуации; поэтому *тут* неестественно в контексте перечислений рутинных последовательностей действий, а также при описании

ситуаций, естественно вытекающих из предшествующих им; в этом смысле *тут* противопоставлено *потом, после этого, вскоре*:

(10а) *Он включил компьютер, потом <и после этого> подсоединил его к проектору*

(10б) ??*Он включил компьютер, и тут подсоединил его к проектору*

(10в) *Он включил компьютер, и тут произошло короткое замыкание*

Ср. также следующую тройку примеров:

(11а) *Он сделал ей предложение, и вскоре она вышла за него замуж*

(11б) ??*Он сделал ей предложение, и тут она вышла за него замуж*

(11в) *Он сделал ей предложение, и тут она сказала ему все, что о нем думает*

Поскольку компонент ‘неожиданность’ прагматически выделен, *тут* в этом значении часто употребляется в комбинации с контрастивно-выделительной частицей — *то*:

(12) *Он сделал ей предложение в четвертый раз, и тут-то она наконец согласилась*

Таким образом, толкование *тут* (назовем *тут* в этом значении *тут*[1]) можно сформулировать следующим образом:

(13) *A1, тут A2* ‘Неожиданная для говорящего ситуация А2 начала иметь место непосредственно после того, как начала иметь место ситуация А1, и в том же реальном или мысленном пространстве, в котором начала иметь место или имеет ситуация А1’.

Сочетание смыслов неожиданности и непосредственного, т. е. быстрого, следования, дает семантический компонент ‘внезапности’, с которым *тут* объединяет одно семантическое свойство. Если *неожиданным* может быть как наступление, так и ненаступление ситуации, *внезапным* может быть только ее наступление. И наречие *внезапно,* и наречие *тут*, в отличие от наречия *неожиданно*, не могут вводить ненаступившую ситуацию; ср. триаду примеров, где первый пример естественен, а второй и третий — неестественны: *Грабитель прыгнул в машину, но она неожиданно не завелась* vs. **Грабитель прыгнул в машину, но она внезапно не завелась* vs. **Грабитель прыгнул в машину, но тут она не завелась*.

Интересным аналогом *тут*[1] является выделительная частица *вот*, в своем основном значении также являющаяся показателем пространственного

дейксиса[4]; ср. *Его назначили главным менеджером, и вот теперь все вечера он проводит на работе.* Прагматически *вот* — антоним *тут*[1], поскольку выражает естественное с л е д о в а н и е ситуации А2 из ситуации А1, в то время как *тут*[1] указывает на неожиданность ситуации А2 для говорящего:

(14а) *Она его ждала, и вот он пришел*

(14б) ?*Она его ждала, и тут он пришел*

(15а) **Она его совершенно не ждала, и вот он заявился*

(15б) *Она его совершенно не ждала, и тут он заявился*

2.1.2. Отличия *тут*[1] от *тут* в значении близости к моменту речи

Назовем *тут* в значении близости к моменту речи *тут*[2]. Между *тут*[1] и *тут*[2] существуют многочисленные семантические и иные различия.

I. Во-первых, как было сказано выше, *тут*[1] связывает некоторые ситуации (обычно в прошлом, реже в будущем), в то время как *тут*[2] связывает некую ситуацию в прошлом, настоящем или будущем с моментом речи:

тут[1]

(16а) *Я вышла гулять с собакой* [А1], *и тут он позвонил мне на мобильный* [А2] [вскоре после ситуации А1]

(16б) *Я выйду гулять с собакой* [А1], *и тут ты позвонишь мне на мобильный* [А2] [вскоре после ситуации А1]

тут[2]

(17а) *Я тут гуляю с собакой* [А1] [в момент речи]

(17б) *Я тут гуляла с собакой* [А1] [незадолго до момента речи]

(17б) *Я тут буду гулять с собакой* [А1] [вскоре после момента речи]

II. Во-вторых, наречие *тут*[1] семантически двухвалентно (речь идет о двух ситуациях, А1 и А2), в то время как для *тут*[2] вторая валентность является как минимум факультативной; ср. недоговоренность примера (18а) при прагматической самодостаточности примера (18б):

(18а) *И тут он приходит* [*тут*[1]]

[4] Подробно переносные значения *вот* и *вон* рассматриваются в работе [Борисова, Овчинникова 2001].

(18б) *Я тут сижу работаю* [*тут*^{2}]

III. В-третьих, в тех случаях, когда упоминаются обе ситуации (ситуация А1 и ситуация А2), распределение ситуаций во времени у *тут*^{2} и *тут*^{1} противоположное: *тут*^{1} вводит вторую ситуацию, т. е. такую, которая следует за чем-то (и синтаксически относится к глаголу, вводящему вторую ситуацию), а *тут*^{2} вводит первую ситуацию, т. е. такую, за которой что-то следует (и синтаксически относится к глаголу, вводящему первую ситуацию). Ср.:

(19а) *Сдает он отчет* [А1]*, и тут оказывается, что половины документов не хватает* [А2] [*тут*^{1}]

(19б) *Я тут заеду в институт* [А1] *и заберу документы* [А2] [*тут*^{2}]

IV. Между *тут*^{1} и *тут*^{2} существует важное просодическое отличие, которое маркирует их разный грамматический статус. *Тут*^{1} — полноценное сентенциальное наречие, которое, как всякое сентенциальное наречие, стоит в начале предложения и может нести на себе, как и всякий полноправный член предложения, акцентное выделение (например, при логическом ударении):

(19) *Сдает он отчет, и ↓тут оказывается, что половины документов не хватает*

Логическое и акцентное выделение может быть дополнительно подчеркнуто присоединением частицы *–то*:

(20) *Сдает он отчет, и ↓тут-то оказывается, что половины документов не хватает*

Тут^2 — клитическая частица, которая не может нести на себе акцентного выделения; ср. неправильность:

(21а) **Я ↓тут была в Лондоне*

(21б) **↓Тут я была в Лондоне*

Если фраза позволяет акцентное выделение *тут* в подобной позиции, значит, *тут* реализуется в другом значении:

(22а) *Я ↓тут работаю (а не там)* [локативное значение *тут*]

(22б) *↓Тут я работаю (а ↓тут — отдыхаю)* [локативное значение *тут*]

Эта особенность *тут*^{2} связана с особым статусом этой единицы: мы предполагаем, что она находится на пути к грамматикализации в качестве особого

показателя актуальности ситуации и ее близости к моменту речи (подробнее особенности *тут* в качестве показателя близости к моменту речи рассматриваются в следующем разделе, в сравнении с наречием *сейчас*). Утрата фонетической самостоятельности и развитие клитичности — один из показателей этой грамматикализации, наряду с частеречными изменениями (превращение из наречия в частицу) и семантическими сдвигами (утрата локативной семантики).

Как видно из вышеприведенных примеров, *тут* может стоять как в позиции проклитики, так и в позиции энклитики, однако чаще *тут* реализуется как энклитика и занимает в предложении вакернагелевсую позицию. В заключение данного раздела приведем контрастный пример на *тут*^{1} и *тут*^{2}, из которого видно, что перестановка *тут* из позиции сентенциального наречия в вакернагелевскую позицию дает смену значения:

(23а) *Тут я вышел во двор* [*тут*^{1}, после наступления какой-то другой ситуации]

(23а) *Я тут вышел во двор* [*тут*^{2}, после этого наступит какая-то другая ситуация]

Соответственно этим двум значениям может быть реконструирован естественный контекст для этих двух *тут*:

(24а) *…и тут я вышел во двор* [*тут*^{1}]

(24а) *Я тут вышел во двор и…* [*тут*^{2}]

3. *Тут и сейчас*

В предыдущих разделах было показано, что *тут* в значении близости к моменту речи обладает рядом особенностей, отличающих его от других временных значений *тут*. В этом разделе предлагается ответить на вопрос о том, чем *тут* отличается от основного дейктического показателя близости к моменту речи — наречия *сейчас*, — и, таким образом, зачем русскому языку еще одна единица с подобным значением.

О значениях *сейчас* и *теперь* подробно писал И. А. Мельчук [Мельчук 1995: 55–75], а также Е. У. Урысон [Урысон 2004:1009–113]. В данной статье *теперь* не рассматривается, поскольку обладает собственной семантической спецификой. Особенность *теперь* по сравнению с более нейтральным наречием *сейчас* состоит в наличии сопоставления настоящего положения вещей с тем, что было в прошлом [Мельчук 1995:72, Урысон 2004: 1011]; ср.:

(25) *Я теперь ему в глаза не смогу смотреть* [после того, что произошло до момента речи]

Тут рассматривается в сопоставлении с нейтральным синонимом *сейчас*, значение которого описывается в цитируемых работах следующим образом:

1) три значения — ‘в момент речи’, ‘в ближайший момент после момента речи’, ‘в ближайший момент перед моментом речи’ [Мельчук 1995: 59];

2) одно синтетическое значение ‘в момент речи или в интервал времени, включающий момент речи’ [Урысон 2004: 1009].

Для наших целей разграничение между употреблениями и значениями не столь существенно, и, поскольку и *сейчас*, и *тут* могут обозначать как момент речи, так и моменты до и после момента речи, мы будем рассматривать эти три возможных реализации как три употребления в рамках единого значения.

I. Первое различие между *тут*[2] и *сейчас* касается статуса этих единиц: как и *тут*[1], *сейчас* является полноценной лексической и просодической единицей, в то время как *тут*[2] *имеет полуграмматикализованный статус. Ниже будет рассмотрен вопрос о том, что именно, по сравнению с сейчас*, выражает эта полуграмматикализованная единица. Здесь ограничимся замечанием, что *тут*[2], в отличие от *сейчас*, не допускает ни логического (или фразового) ударения с соответствующим просодическим выделением, ни какого-либо другого помещения в фокус внимания; ср. контрастные примеры:

(25а) — *Когда ты это будешь делать? — Сейчас.*

(25б) — *Когда ты это будешь делать? — *Тут.*

Ср. также

(26а) *И не просто сейчас, а прямо сейчас, движется-крутится просто непонятное количество разных молекул* (Е. Гришковец).

(26б) **И не просто тут, а прямо тут, движется-крутится просто непонятное количество разных молекул* [невозможно с временн_ой интерпретацией].

Как мы видим из этих примеров, помещение *тут*[2] в фокус внимания неизбежно приводит либо к аномалии, либо к локативной интерпретации: единица *тут* в значении близости к моменту речи возможна только в качестве грамматикализованного показателя. Интересно, что *тут*[2] часто употребляется в сочетании с *сейчас* (*Я тут сейчас отчет пишу, и мне нужны показатели за сентябрь*), что было бы невозможно, если бы эти две единицы обладали одинаковым статусом, как невозможно сочетание *сейчас* с его близким синонимом *в данный момент* (**Я сейчас в данный момент отчет пишу, и мне нужны*

показатели за сентябрь). Ср. также литературные примеры на сочетание *сейчас* и *тут*^{2}, где *тут*^{2} явно используется не в локативном значении и, на первый взгляд, дублирует значение *сейчас*:

(27) *Миля, — сказал он, — тут сейчас будет звонить опять Потоцкая, так я у Якова Абрамовича в триста пятидесятой* (Ю. Домбровский)

II. Проанализируем теперь смысловые различия между *сейчас* и *тут*^{2} в качестве показателей близости к моменту речи. Сначала рассмотрим те (более многочисленные) группы примеров, где возможно употребление *сейчас*, но не *тут*^{2}. Отметим предварительно, что все примеры, где наречие *сейчас* коммуникативно и просодически выделено, т. е. является ремой или контрастной темой, автоматически невозможны для *тут* и не будут рассматриваться. Это примеры типа *↓Сейчас я живу в Америке*; *Поговори об этом ↓сейчас, не откладывай*; *Если бы он занялся наукой, мог бы ↓сейчас быть академиком.* Таким образом, примеры, которые приводятся ниже, это примеры, которые коммуникативно и просодически являются для *тут*^{2} допустимыми.

(28а) *В Антарктиде сейчас морозы за сорок*

(28б) **В Антарктиде тут морозы за сорок*

Пара (28а) и (28б) наводит на мысль, что для *тут*^{2} невозможны эксплицитно локативные контексты. Однако это не так; ср. (29а) и (29б):

(29а) *В Москве сейчас опять жара под сорок*

(29б) *В Москве тут опять жара под сорок*

Можно предположить, что *тут*^{2}, в силу остаточной пространственной семантики, действительно накладывает ограничения на употребление в локативных контекстах, а именно, как и пространственное *тут*, требует, чтобы говорящий находился в том месте, о котором идет речь. Однако и это не является вполне верным; ср.:

(30) *В Финляндии тут в очередной раз повысили цену на алкоголь, и все стали ездить за водкой к нам*

Таким образом, пространство, о котором идет речь, может физически не совпадать с тем, в котором находится говорящий, но м ы с л и т ь с я должно как часть пространства говорящего (в этом смысле *тут*^{2} не отличается от *тут*^{1}). Тот факт, что говорящий хотя бы ментально находится в том пространстве, о котором он делает высказывание, подтверждается также невозможностью употребления *тут* в применении сразу к двум пространствам в пределах одного высказывания (т. к. говорящий не может находиться в двух пространствах

одновременно, по-видимому даже ментально); ср. неправильность **В Москве тут лето, а в Кейптауне тут зима.*

Можно, конечно, предположить также, что в данном случае мы сталкиваемся не с временным, а с размытым локативным значением *тут*, похожем на *там*; ср. *Там в Финляндии в очередной раз повысили цену на алкоголь…* Однако это предположение легко опровергается; ср. невозможность употребления *тут*[2], если речь идет о давнопрошедших событиях, даже если событие имело место там, где находится говорящий:

(31) **В России тут в 19 веке была война с Наполеоном*

Правильные фразы типа (32) не являются контрпримерами, так как в них реализуется локативное значение *тут*, и *тут* является сентенциальным наречием (в отличие от частицы *тут*[2]).

(32) *Тут, в России, в 19 веке была война с Наполеоном*

Однако если событие произошло относительно недавно, то фразы типа (31) становятся возможными:

(33) *В России тут прошла волна политических протестов*

Итак, сформулируем важное различие между *сейчас* и *тут*[2]. *Сейчас* не накладывает никаких ограничений на местонахождение говорящего, но только на время события, о котором идет речь. *Тут*[2] требует, чтобы не только временной промежуток, о котором идет речь, включал в себя момент речи, т. е. мыслился говорящим как тот, в котором он сейчас живет, но и чтобы п р о с т р а н с т в о , в котором происходит/ происходило/ произойдет событие, мыслилось говорящим как включающее его.

III. Рассмотрим теперь ситуации, когда возможно употребление *тут*[2], но не *сейчас*. Ср. следующие пары:

(34а) **Я сейчас в прошлом году была в Лондоне*

(34б) *Я тут в прошлом году была в Лондоне*

(35а) **Я сейчас в будущем году буду в Лондоне*

(35б) *Я тут в будущем году буду в Лондоне*

При этом возможны фразы с *сейчас*, где это наречие указывает на моменты до и после момента речи.

(36а) *Я сейчас была в Лондоне*

(36б) *Я сейчас буду в Лондоне*

Таким образом, как видно из этих примеров, *тут*^{2} допускает большую удаленность момента события от момента речи, нежели *сейчас*. Более того, для *сейчас*, по-видимому, не возможны никакие уточнения касательно того, когда именно произошло событие, даже если промежуток времени, отделяющий его от момента речи, очень мал:

(37а) **Я сейчас пять минут назад его видела*

(37б) **Я сейчас через пять минут буду с ним говорить*

Для *тут*^{2}, как мы видим, это возможно, даже если временной промежуток объективно не так уж и мал. Однако примеры типа (38а) и (38б) уже невозможны:

(38а) **Я тут тридцать лет назад была в Лондоне*

(38б) **Я тут через тридцать лет выплачу долг за квартиру*

Как же можно сформулировать временные ограничения на употребление *тут*^{2}?

По-видимому *тут*^{2} относится либо к моменту речи, либо к такому моменту до или после момента речи, когда то, что произошло/ произойдет, может быть актуально для настоящего:

(39а) *Я тут год назад была в Лондоне и купила себе кучу одежды* [имеется в виду, что я пользуюсь ею и сейчас, и поэтому у меня есть, в чем пойти в гости/не нужно покупать новую и т. п.]

(39б) *Я тут через год выплачу кредит за квартиру* [имеется в виду, что вскоре мое материальное положение улучшится — например, я смогу купить машину/ дать кому-то в долг/ пожертвовать на благотворительность и т. п.]

В этом смысле значение *тут*^{2} отчасти сближается со значениями некоторых видовременных глагольных форм в европейских языках, например английской формы Present Perfect (хотя *тут*^{2} в равной степени ориентировано на связь настоящего с прошлым и с будущим, а Present Perfect — только на связь прошлого с настоящим), французского passé recent и futur proche. При этом *тут*^{2}, безусловно, еще не достигло окончательной грамматикализации и его значение гораздо более конкретно, чем значение граммем.

IV. Рассмотрим последнее отличие *тут*^{2} от *сейчас*. В нижеследующих фразах *сейчас* является невозможным или звучит менее уместно, нежели *тут*^{2}.

(40а) *Мне тут влепили двойку по геометрии* (40б) ?Мне сейчас влепили двойку по геометрии

(41а) *Мы тут полагаем, что пора уже действовать*

(41б) **Мы сейчас полагаем, что пора уже действовать*

(42а) *Я тут кое-что о нем знаю, но вам не скажу*

(42б) **Я сейчас кое-что о нем знаю, но вам не скажу*

Пример (40б) не идеален, но не невозможен: если говорящий хочет сообщить именно и только о времени получения двойки и о том, что это событие произошло непосредственно перед моментом речи, то пример (40б) прагматически адекватен. Что касается примеров (41б) и (42б), то они аномальны, потому что мнение (тот его вид, который выражается глаголом *полагать* — устойчивое мнение) и особенно знание по своей природе несовместимы с идеей существования исключительно в момент речи, которая вводится наречием *сейчас* в сочетании с глаголом в настоящем времени. У знания есть начало (когда человек что-то узнает), но конец не предусматривается. Устойчивое мнение, выражаемое глаголом *полагать* (в отличие от *думать*) также не предполагает быстрой смены. Однако в сочетании с *тут*^{2} и *полагать*, и *знать* возможны, из чего можно сделать вывод, что *тут*^{2} выражает нечто помимо идеи настоящего момента. Уже было сказано, что *тут*^{2} также предполагает местонахождение говорящего в том же мысленном пространстве, о котором идет речь. По-видимому, *тут*^{2} имеет синтетическое значение актуальности, в котором объединены разные идеи: период времени и пространство, в котором говорящий себя мыслит, а также ситуация, участником или непосредственным наблюдателем которой он является. Кроме того, как было показано в разделе 2.1.2, *тут*^{2} иногда предполагает, что за ситуацией A1, о которой сообщает говорящий, следует какаято другая ситуация A2, обычно вытекающая из той, которая имеет место: *Я тут летом собираюсь в поход* [A1], *и у всех спрашиваю, нет ли палатки* [A2]. Этим *тут*^{2} также отличается от чисто временного *сейчас*.

Таким образом, идея времени встроена в значение *тут*^{2}, однако не является в нем единственной. В этом смысле *тут*^{2} и *тут*^{1} сближаются: обе единицы предполагают некий репортаж, некую эвиденциальность — в противоположность объективному нарративу. Говорящий сообщает о том, к чему он имеет прямой непосредственный доступ — при этом *тут*^{2} предполагает, что ситуация разворачивается в момент речи или в такой момент времени до или после момента речи, когда она еще сохраняет актуальность для настоящего и ее связь с настоящим не утрачена, а *тут*^{1} обычно сообщает о событиях в прошлом и никакой связи с настоящим не предполагает.

Поскольку *тут*^{2}, в отличие от *сейчас*, имеет не чисто временное значение, для него возможны сочетания с обстоятельствами времени, невозможные для чисто временного *сейчас*; ср.:

(43а) *Я тут летом в поход собираюсь*

(43б) *Я тут в октябре была в Лондоне*

(43в) *Я тут в пять должна уходить* (44а) **Я сейчас летом в поход собираюсь*

(44б) **Я сейчас в октябре была в Лондоне*

(44в) **Я сейчас в пять должна уходить*

В заключение сформулируем значение *тут*[2]:

(45) *Тут A1, A2* ‘Имеет место ситуация A1, участником или наблюдателем которой часто является говорящий, которая разворачивается в том же реальном или мысленном пространстве, где находится говорящий, в период времени, который включает в себя настоящий момент и такие моменты до и после настоящего момента, когда ситуация A1 сохраняет связь с настоящим моментом’.

То, что говорящий часто является участником или наблюдателем ситуации, подтверждается языковыми фактами. Во-первых, в тех случаях, когда говорящий не является центральным участником ситуации, т. е. ее субъектом, он обычно выступает либо в роли наблюдателя, либо в роли периферийного участника; ср.

(46) *Он тут кричит на своего ребенка* [говорящий является непосредственным наблюдателем]

(47) *Петька тут ногу сломал* [говорящий является косвенным участником ситуации; такая фраза возможна, если Петька входит в личную сферу говорящего, и это событие каким-то образом говорящего затрагивает — например, он должен везти Петьку в больницу]

Роль говорящего в ситуации, предполагаемая частицей *тут*, особенно ярко проявляется в сочетаемости *тут* с ментальными предикатами типа *подумать, узнать, решить, понять*. Так, в основном корпусе НКРЯ на 53 употребления *тут* с субъектом первого лица (типа *Я тут подумал, Я тут решил*) приходится всего лишь четыре употребления *тут* с субъектом второго лица (типа *Что вы тут вздумали*) и одно употребление с субъектом третьего лица (типа *Он тут понял, что она его предала, и решил отомстить*). Таким образом, тот факт, что большинство примеров на *тут*[2], использующихся в данной статье, имеют субъект первого лица, не является случайной аберрацией, а отражает мотивированную семантикой грамматическую сочетаемость.

Литература

1. *Е. Г. Борисова, Т. Е. Овчинникова*. Параметр близости в метафорическом пространстве // Компьютерная лингвистика и интеллектуальные технологии /По материалам ежегодной Международной конференции «ДИАЛОГ» (2011), Вып. 10. М.: РГГУ, 2011. с. 153–158.
2. *И. А. Мельчук*. ‘СЕЙЧАС’ и ‘ТЕПЕРЬ’ в современном русском языке // И. А. Мельчук, Русский язык в модели «Смысл — Текст». Москва — Вена: Школа «Языки русской культуры», Венский славистический альманах, 1995. — С. 55–79.
3. *Е. У. Урысон*. Синонимический ряд СЕЙЧАС 1 // Новый объяснительный словарь синонимов русского языка. Издание второе, исправленное и дополненное. Под общим руководством академика Ю. Д. Апресяна. Ю. Д. Апресян, В. Ю. Апресян, Е. Э. Бабаева, О. Ю. Богуславская, И. В. Галактионова, С. А. Григорьева, Б. Л. Иомдин, Т. В. Крылова, И. Б. Левонтина, А. В. Птенцова, А. В. Санников, Е. В. Урысон. М., 2004. с. 1009–1013.
4. *T. Fretheim* (April 2008). Reference to the near past by means of the Norwegian “spatial” term ‘her’ (here). Unpublished paper presented at Space and Time across Language, Disciplines and Cultures conference, Camridge, UK.

RUSSIAN INTONATION CORPUS: A PRELIMINARY REPORT

Arkhipov A. V. (sarkipo@yandex.ru),
Zakharov L. M. (leon@philol.msu.ru),
Krivnova O. F. (okri@philol.msu.ru),
Kodzasov S. V. (sankod@yandex.ru),
Lebedev A. A. (andrettile@inbox.ru)

M. V. Lomonosov Moscow State University, Moscow, Russia

The paper presents Russian Intonation Corpus (RINCO), the multimedia speech corpus designed for the study of prosody in Russian, its history, purpose, and functionality. RINCO started as a series of MS Access databases developed in 2004–2009 by a research team at the philological faculty of Moscow State University led by S. V. Kodzasov. The databases contained descriptions of most important prosodic characteristics (several tonal features, tempo, loudness, phonation types) for quasi-natural dialog utterances, as well as for a selection of spontaneous and prepared spoken narrative texts of different genres. These descriptions have been transferred to a multi-layer time-aligned markup format, suitable for local use in ELAN annotator and for online browsing, playback and search using the LAT server platform. The corpus in its current state should be fully operational at http://languedoc.philol.msu.ru/rinco by the end of 2012. Various possibilities of enhancing the corpus are discussed, as well as some findings with a sample query for "IK-5" intonational construction.
The project is currently supported by RFH grant #11-04-12042.

Keywords: multimedia corpus, intonation, prosody, Russian language

1. Project overview

The Russian Intonation Corpus (RINCO) originated in 2004 as a project for creating an array of dialogical utterances provided with complex prosodic markup. The scale of the data and the nature of the markup were adapted for the purposes of linguistic investigations rather than direct use in speech technologies. In the following three years (2004–2006), the team led by Sandro V. Kodzasov developed three MS Access databases, each devoted to a specific type of utterances (questions, imperatives, statements). Another three-year project was undertaken to analyze monological spoken texts of different genres or their fragments: TV news programs, literary and traditional prose reading, and real life stories (2007–2009). These two projects were supported by RFH grants # 04-04-12027v and # 07-04-12160v and presented in a number of previous publications in DIALOG proceedings (see e. g. Kodzasov et al. 2005, 2006, 2008).

The goal of the current project is to bring the previously collected data to a new level of efficiency and accessibility. The main issues addressed are (1) open web access and feedback, (2) time-alignment of the prosodic features' markup, and (3) improved search

capabilities. In order to attain (1)–(3), the MS Access databases were ported to a new platform, consisting of ELAN (for local use) and the Language Archiving Technologies (LAT) software suite (for online use), both developed at Max-Planck Institute for Psycholinguistics, Nijmegen (Netherlands). At the same time, we tried to enhance the markup with some new tiers of information, as well as to develop documentation for a wider audience (in progress).

The current project, scheduled to finish in 2012, will lay ground for a full-scale multimedia speech corpus of Russian for intonation studies which can be further enhanced in several directions, including dataset expansion, introducing new kinds of markup (e. g. alternative transcription systems), improving metadata, navigation and help resources, user interface and export options.

2. Corpus presentation

2.1. Main features

The main features that make RINCO special compared to other existing corpora are the following:

- RINCO is centered around the recordings of speech, not around its transcripts (unlike e. g. the spoken subcorpus of the Russian National Corpus).
- The main motivation behind RINCO was to illustrate the prosodic variation in spoken Russian and thus to cover the most diverse utterance types, rather than to construct a weighted representative corpus.
- For the same reason, the data for the first (dialogical) subcorpus were recorded in studio rather than taken from spontaneous speech recordings, since observation of natural speech would necessitate huge volumes of data to cover all the desired utterance types.
- The prosodic markup used in RINCO follows the combinatory model of prosodic features by Sandro V. Kodzasov [Kodzasov 1996].
- The target audience includes primarily researchers in phonetics & phonology as well as in semantics & pragmatics. RINCO was not designed specifically for use in ASR or TTS systems; however, we hope that fruitful partnership is possible with experts in speech technologies. Also, RINCO can also be useful for students and teachers of Russian (both as second and first language, e. g. as an aid in mastering intonation for people with hearing impairments).

2.2. Data

The first subcorpus, “Intonation of Russian Dialogue” (IRD), includes around 1,000 utterances of three major types (questions, imperatives, and statements), with around 300–400 utterances for each type, totalling ca. 90 minutes recording time. These are utterances pronounced in isolation or in “cue-response” pairs. The textual examples that were recorded originate from authors’ observations and recordings or published linguistic works, partially adapted or constructed to cover most

important semantic, pragmatic, and grammatical types. The speakers who participated in recording sessions were Moscow inhabitants with higher education, all native speakers of Standard Russian, both men and women, aged from 20 to 70.

The second subcorpus included (fragments of) monological discourse of different genres: live TV news programs, literary and traditional narrative texts (novels, folk tales) recorded by professional readers/actors, and real life stories told by several speakers. The overall time of the monological subcorpus is around 90–100 minutes.

The primary data was principally recorded in Linear PCM .wav format (except for the database on questions which was first recorded on magnetic tape and then digitized).

2.3. Markup

The linguistic annotations provided for the speech samples include multiple layers (tiers) of markup, time-aligned to the speech signal where appropriate. The tiers used in the IRD subcorpus include:

- *Text* — Text of the utterance in orthographic representation.
- *Transcribe* — Canonical phonetic transcription as generated by the transcriber component of the NRS speech synthesis package. This is not (yet) time-aligned or manually checked.
- *Voice* — Male or female.
- *Int phrase* — Same as Text, segmented into intonation phrases.
- *Words* — Same as Text, segmented into words.
- *Accent* — Local tonal accents (includes localization and directions of pitch movement, interval of pitch change and some non-tonal prosodic accents (e. g. loudness or duration)).
- *Tone* — All non-automatic pitch movements.
- *Pause* — Pauses longer than 0.2 sec
- *Loudness, Tempo, Register, Phonation* — Integral prosodies.
- *Type 1*, *Type 2* — Utterance type according to the complex semantic and grammatical classification developed by project members. Only the highest and the lowest levels of classification are included in the annotation file itself; however, all the levels of classification are represented in (browsable and searchable) metadata which form the tree structure of the corpus (see below).
- *Comments*
- *ID* — Internal code of the utterance.

The monological subcorpus uses the same tiers except for the classification tiers (Type 1, Type 2).

2.4. Setup for local use

The markup was implemented in ELAN multimedia annotator [Wittenburg et al. 2006]. ELAN was initially designed for the purposes of studying sign language and

is also widely used in various endangered language documentation projects. ELAN supports an arbitrary number of text annotation tiers time-aligned to the sound (and/or up to four video files). ELAN can also display pitch graphs generated by the Praat phonetic analyzer (*.PitchTier). Annotation tiers can be organized in tree-like hierarchy with several possible types of relation between them.

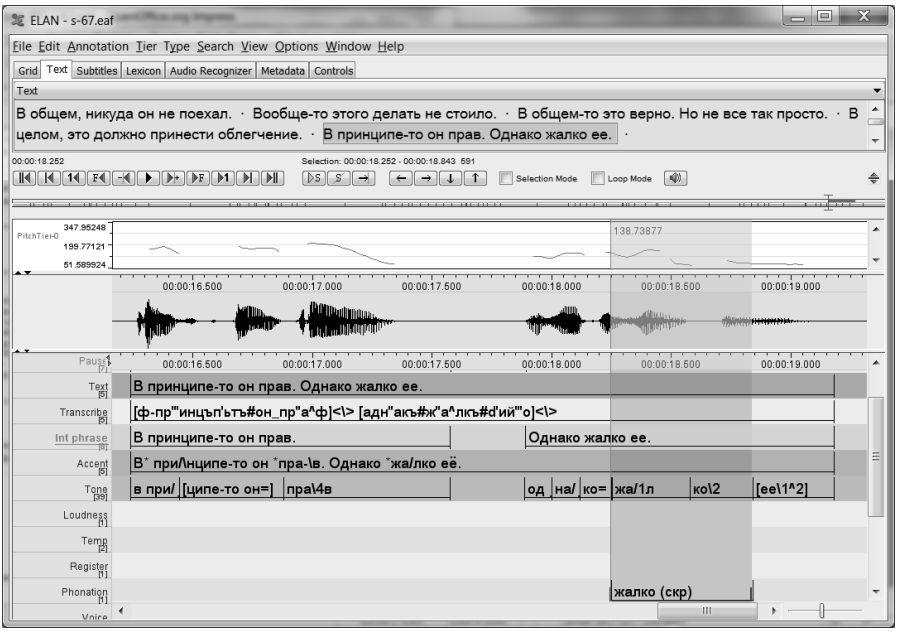


Fig. 1. Sample ELAN window

ELAN can be used to create, edit, and browse the annotated media files, as well as to run complex search queries within a single file and across large collections of files. The search engine allows for searching using regular expressions, searching by specifying various constraints on structural and temporal relations between annotations within the same tier or across multiple tiers. Search results can then be explored in ELAN or exported e. g. as CSV tables.

2.5. Online implementation

The online version of the corpus uses the Language Archiving Technologies (LAT) platform which includes several pieces of software, as the IMDI browser, serving as the main interface to the corpus, and the ANNEX utility used for online access to the annotation files, offering much of the ELAN functionality to explore (not edit) ELAN files (and some other kinds of text annotation formats) online, except for online display of pitch graphs. In particular, ANNEX allows for playing back the media

fragment corresponding to the selected time interval or annotation unit. A separate utility, TROVA, offers an interface to the search engine (same as used with ELAN).

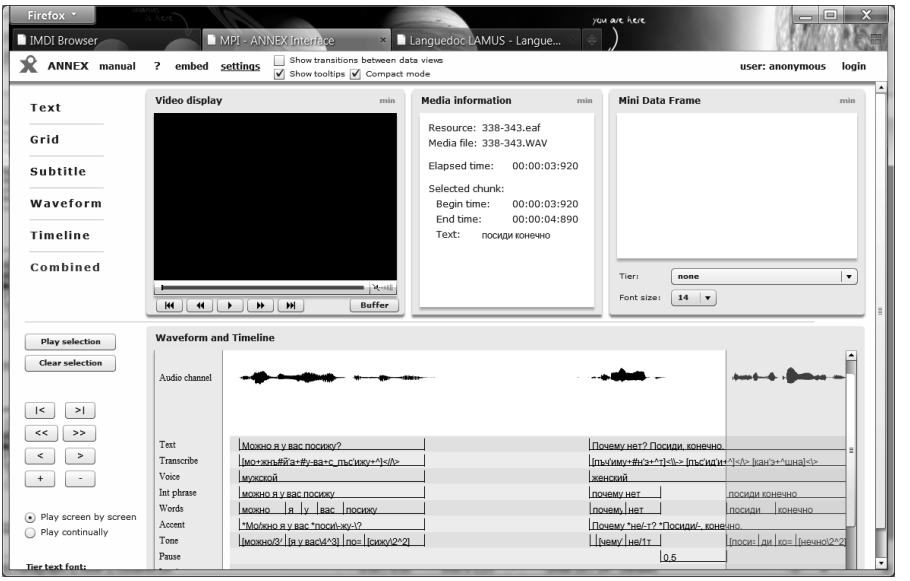


Fig. 2. Russian Intonation Corpus online: ANNEX viewer

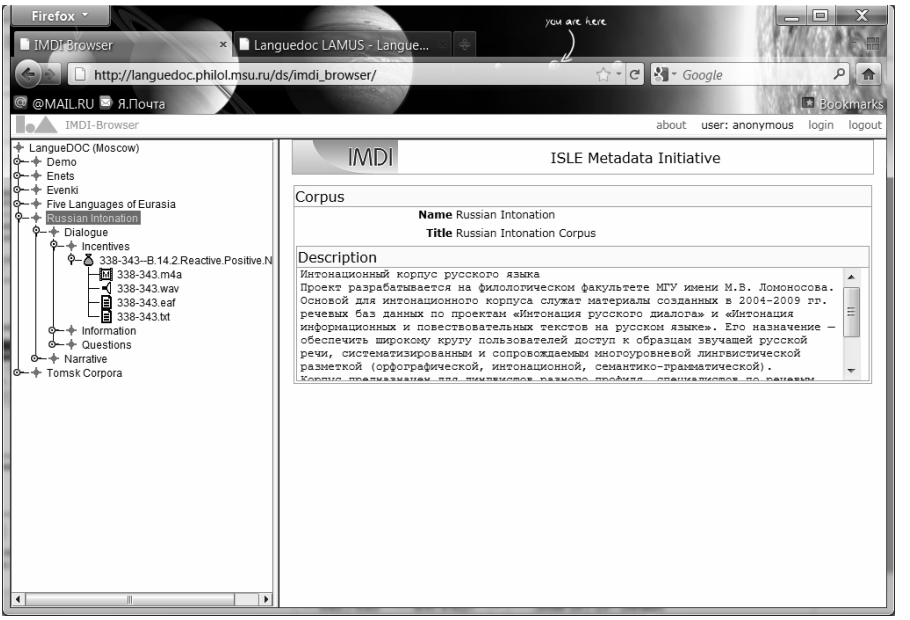


Fig. 3. Russian Intonation Corpus online: IMDI metadata browser

The IMDI browser presents the corpus structure which is calculated based on metadata descriptions of the corpus files. In the case of RINCO, the metadata include information about the project, the speakers/sources of data, media files, the types of utterances (for the IRD subcorpus) according to the adopted classification, the textual content (for the monological subcorpus), and on the access rights. The metadata can also be searched to form a subcorpus (which can be further searched by content).

The IMDI metadata are by principle visible to all users; the corpus creators can further specify access rights to particular files or file types. RINCO offers online access for all users to browse the annotations and play back media files, which means that ELAN annotation files and Praat pitch graphs can be freely downloaded for local use. Access to sound files will be available on demand.

3. An illustration: The uses of "IK-5"

Let us consider a simple search example and look for the tonal "hat contour," traditionally termed "IK-5" ("intonational construction 5") [Bryzgunova 1977: 78; 1980: 116]. A standard IK-5 is usually described as having two nuclei with rising tone on the first one, falling tone on the last one and a level tone (no significant pitch change) between them. This can be searched for as a sequence of "/-" and "-\" tonal accents in the Accents tier, or as a sequence of "/", "=" and "\" movements (in adjacent annotations) in the Tone tier (see Fig. 4).

IK-5 has been initially described as typical for a particular utterance type, namely for exclamatives of admiration, especially with WH-words (hereafter the rise and fall are marked with /- and -\ after the vowel, respectively; the dash — signals the absence of an automatic fall after rising, as opposed to the unmarked / and \ tonal accents):

(1) *Како/-е у-\тро! 'What a (wonderful) morning!' (from [Bryzgunova 1980: 116])*

(2) *За/-меча-\тельно! 'Wonderful!' (from [Bryzgunova 1980: 116])*

However, subsequent studies emphasized that these exclamatives would normally call for a particular phonation type (breathy voice, sometimes also lax voice, and slower tempo) [Kodzasov 1996/2009: 21]. On the other hand, the same tonal configuration with modal voice was shown to be typical of WH-questions [Nikolaeva 1977]. We do indeed find both of these kinds of uses in RINCO, as illustrated in (3) and (4); utterance ID in the corpus is given in square brackets.

(3) *У неё прекра/-сный вку-\с! 'She (has) an excellent taste!' [st79f]*

(4) *Покажите, пожалуйста, ка/-к им по-\льзоваться? '(Could you) please show how to use it?' [i95m]*

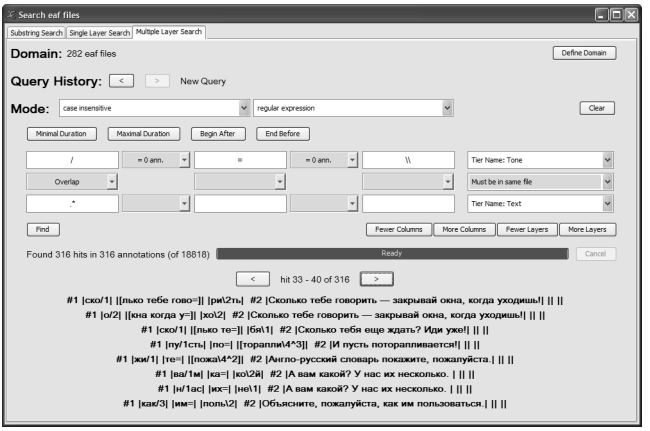


Fig. 4. A sample multi-tier search window

Among questions, one can mention highly emotional questions with astonishment or indignation, which usually demonstrate a greater range of pitch change (above 7–8 semitones):

(5) *Неуже/-ли он об этом не зна-\л? 'Did he really not know about it?' [q122m]*

(6) *Отк/-уда мне зна-\ть? 'How would I know?' [q195f]*

(7) *Заче/-м мне было его бра-\ть? 'Why would I take it?'[q215m]*

(8) *А что/- я такого сказа-\л? 'What did I say?' [q216m]*

However, a significant number of examples falls outside of these classes. We find, for instance, a number of similar configurations in requests, especially in a more formal setting:

(9) *Бу/-дьте любе-\зны, говорите громче! 'Please speak louder!' [i5Af]*

(10) *Прошу/- сади-\ться. 'Please be seated.' [i152m]*

(11) *Англо-русский словарь покажи/-те, пожа-\луйста. '(Could you) please show me the English-Russian dictionary.' [i94Am]*

(12) They can also appear on (cooperative) forms of address, corroborating what had been reported in [Kodzasov 1998/2009: 163]:

(13) *Пё/-тр Ива-\нович, за/-втрака-\ть! 'Petr Ivanovich, (please come) to have breakfast!' [i353f]*

Another use of 'hat contour' is in replies like (13):

(14) *{Полкило творога, пожалуйста.} — А о/-н не очень све-\жий. [i57Af]*
'{Half a kilo cottage cheese, please.} — (But) it is not very fresh.

In general, the combination of rise on Topic and fall on Focus is quite common in statements. However, sustaining high level tone after the rise seems to be correlated with statements used as an indirect negative reply; cf. the following pair:

(15) *Ивано/в заболе\л. 'Ivanov fell ill.'*

(16) *{Иванова можно к телефону?} — А Ивано/-в заболе-\л.*
'{Can I talk to Ivanov(, please)?} — Ivanov has fallen ill.' [Kodzasov 1996/2009: 22]

Apart from the rise-level-fall configuration, one can also observe a comparable number of the opposite contour, fall-level-rise. It is particularly common in questions with Topic preceding Focus, in which case the Topic adapts inversely to the rise on the Focus [Kodzasov 1996/2009: 22]:

(17) *Посу\-ду помо-/ешь? 'Will you wash the dishes?' [i115f]*

(18) *Ча\-йник не поста-/вишь? 'Will you turn on the kettle?' [i117.2m]*

(19) Cf.: *Ты не поставишь чайник? 'Will you turn on the kettle?' [i117.1f]*

(20) *Васю пригласить? 'Shall I invite Vasia?' [i380Af]*

When a question word bearing raising tone precedes the Topic (inversely marked with falling tone), a rise may appear on the utterance-final syllable to keep marking it as a question:

(21) *Что/ до\ма/? 'What (news) at home?' [q143f]*

With this quick glance of IK-5 and related tonal contours, we hope to have demonstrated the usefulness of the possibility to search for specific prosodic patterns, which can serve teachers and students of Russian intonation as well as researchers.

4. Current state and outlook

The current project is scheduled to finish by the end of 2012, when all the data should have been uploaded to the project server (http://languedoc.philol.msu.ru/rinco). However this will only be the beginning of work on the large-scale corpus of Russian intonation. Several directions of future work can be foreseen.

Most naturally, the range of primary data should be expanded, possibly with recordings of natural (spontaneous) speech of different genres, including natural dialogs.

The prosodic markup can be further enhanced in several ways, namely adapting the automatic phonetic transcription to the real pronunciation; using the phonetic representation as the basis for prosodic markup (instead of the orthographic representation); introducing alternative prosodic transcription schemes — e. g. the ToRI system ([Ode 2003]) and the widely used traditional Russian notation by Elena A. Bryzgunova [Bryzgunova 1977]. The latter will require collaboration by experts in corresponding transcription systems, which can be easily implemented thanks to the customizable markup format of ELAN.

Finally, user experience may be improved by providing detailed metadata descriptions, various navigation aids and help resources, developing the user interface(s) and additional export options.

References

1. *Bryzgunova, E. A.* Sounds and intonation of Russian speech [Zvuki i intonatsiia russkoi rechi]. Moscow, 1977.
2. *Bryzgunova, E. A.* Intonation [Intonatsiia], in Russian Grammar [Russkaia grammatika], v. 1. Moscow, 1980. P. 96–120.
3. *Kodzasov S. V.* Combinatory model of the sentence prosody. [Kombinatornaia model' frazovoi prosodii], in Prosodic structure of Russian speech [Prosodicheskii stroi russkoi rechi]. Moscow, 1996. Reprinted in Kodzasov 2009.
4. *Kodzasov S. V.* The prosody of forms of address. [Prosodiia obrashchenii], in Language: variability and stability [Iazyk: izmenchivost' i postoianstvo]. Moscow, 1998. Reprinted in Kodzasov 2009.
5. *Kodzasov S. V.* Studies in Russian prosody [Issledovaniia v oblasti russkoi prosodii]. Moscow, 2009.
6. *Kodzasov S. V., Bonch-Osmolovskaia A. A., Zakharov L. M., Kobozeva I. M., Krivnova O. F.* Database "Intonation of Russian Dialogue": interrogative utterances [Baza dannykh "Intonatsiia russkogo dialoga": voprositel'nye repliki]. Trudy Mezhdunarodnoi Konferentsii "Dialog 2005" [Proceedings of the International Conference "Dialog 2005"]. Moscow, 2005. P. 245–249.
7. *Kodzasov S. V., Arkhipov A. V., Bonch-Osmolovskaia A. A., Zakharov L. M., Krivnova O. F.* Database "Intonation of Russian Dialogue": imperative utterances [Baza dannykh "Intonatsiia russkogo dialoga": pobuditel'nye repliki]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2006". [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2006"] Bekasovo, 2006. Moscow: RSUH, 2006. P. 236–242.
8. *Kodzasov S. V., Arkhipov A. V., Zakharov L. M., Krivnova O. F.* Database "Intonation of Russian Informational Texts" [Baza dannykh "Intonatsiia russkix informatsionnykh tekstov|"]. Komp'iuternaia Lingvistika i Intellektual'nye

Tekhnologii: Po materialam ezhegodnoi mezhdunarodnoi konferentsii “Dialog 2008”. [Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”] Bekasovo, 2008. Issue 7 (14). Moscow: RSUH, 2008. C. 206–208.

9. *Nikolaeva T. M.* Sentence intonation of Slavic languages [Frazovaia intonatsiia slavianskikh iazikov]. Moscow, 1977.
10. *Odé, C.* Developing a Transcription of Russian Intonation (ToRI). Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, August 3–9, 2003. P. 3097–3100.
11. *Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.* ELAN: a Professional Framework for Multimodality Research. Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation. 2006. [http://www.lat-mpi.eu/tools/elan/]

ПРОЕКТ ЧАСТОТНОГО СЛОВАРЯ РУССКИХ ИДИОМ

Баранов А. Н. (baranov_anatoly@hotmail.con),
Вознесенская М. М. (voznes-masha@yandex.com),
Добровольский Д. О. (dm-dbrv@yandex.ru),
Киселева К. Л. (xenkis@mail.ru),
Козеренко А. Д. (akozerenko@gmail.com)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

В русской лексикографии до сих пор не существует частотного словаря идиом. Составление такого словаря осложняется тем, что идиомы встречаются в текстах относительно редко. Предлагается проект создания такого словаря на основе словника «Словаря тезауруса современной русской идиоматики», включающего около 8000 идиом, и четырех корпусов текстов. Для получения репрезентативного корпуса текстов необходимо учесть жанровое своеобразие некоторых идиом и индивидуальные особенности языка писателя. При составлении словника предусмотрено разрешение омонимии. Словарь будет содержать информацию об относительной частотности идиом в корпусе в целом, в отдельных корпусах и в подмножествах корпусов, отражающих разные речевые жанры и типы дискурса, а также о частотности и фразеологической активности отдельных компонентов, входящих в состав идиом.

Ключевые слова: частотный словарь, идиомы, лексика, фразеология

Настоящая работа посвящена обсуждению проекта частотного словаря русских идиом. Поскольку словарей такого типа в лексикографии не существует, подобный проект требует серьезной теоретический и практической подготовки. Однако самые общие принципы создания фразеологического частотного словаря, вероятно, могут быть заимствованы из известных частотных словарей обычной лексики. Рассмотрим существующую практику составления частотных словарей лексики русского языка.

1. Частотные словари русской лексики: принципы составления

Среди существующих частотных словарей современного русского языка следует упомянуть в первую очередь «Частотный словарь русского языка» (1977) под редакцией Л. Н. Засориной, «Статистический словарь языка русской газеты (1990-е годы)», «Частотный словарь современного русского языка

(на материалах Национального корпуса русского языка)» (2009), а также «Частотный словарь современного русского языка» под ред. Л. Лённгрена (1993). Первым современным частотным словарем русского языка следует, по-видимому, считать Частотный словарь (1963) Э. Штейнфельдт, основанный на корпусе советской литературы для школьников объемом 1 млн. словоупотреблений и преследовавший в первую очередь учебные цели. Наиболее известен «Частотный словарь русского языка» (1977) под редакцией Л. Н. Засориной, который до недавнего времени был основным русским частотным словарем. Этот словарь, так же как и «Частотный словарь современного русского языка» (1993) под ред. Л. Лённгрена, опирается на относительно небольшой по современным меркам корпус, насчитывающий около 1 млн. словоупотреблений, в котором представлены тексты с начала XX века и до 1960-х годов. Словарь охватывает четыре речевых жанра: художественная литература, разговорная речь, научные и публицистические тексты, а также деловая речь; за единицу словника словаря принята лексема. Поскольку основная часть публицистических текстов и в значительной мере литературных произведений, вошедших в состав корпуса, были идеологически маркированы, словарь в целом преимущественно отражает официальный русский язык советского периода.

Специализированные статистические словари представляют отдельные жанры: язык художественной литературы (Josselson 1953), науки (Степанова 1970), публицистики (Шайкевич и др. 2008). Так, статистический словарь языка русской газеты (2008) исследует только публицистику 1990-х годов и опирается на несколько корпусов газетных текстов общей численностью более 50 млн. словоупотреблений. Отметим, что в этот словарь в качестве вокабул включены не только отдельные слова, но и фразеологические словосочетания — в основном коллокации, а также прагматемы и некоторые идиомы (ср. включение в словарь таких единиц как *круглый стол, по крайней мере, Филькина грамота*), однако подобных вокабул всего несколько десятков.

Существуют также словари, содержащие статистическую информацию по языку одного писателя — например, Грибоедова (Поляков 1999), Достоевского (Шайкевич и др. 2003), Цветаевой (Белякова и др. 1996). Статистический словарь языка Достоевского (2003) построен на корпусе текстов в 2.9 млн. словоупотреблений, разбитом на подкорпусы трех основных жанров — художественная литература, критика и публицистика, письма. Наиболее частотные лексемы также распределены по периодам творчества автора. Статистическая информация в словаре представлена в виде множества таблиц: таблица частоты употребления лексем по трем жанрам, обратный частотный словарь, распределение некоторых частых лексем художественной литературы по периодам творчества, таблица частоты некоторых бинарных сочетаний, таблицы текстуальных связей некоторых слов, таблицы лексических маркеров персонажей, таблицах грамматических признаков и др. В результате объем полученных статистических данных оказался настолько велик, что авторы приняли решение выпустить словарь в гибридном виде: однотомная книга, показывающая лишь часть таблиц, и сопровождающий ее компакт-диск, содержащий полную информацию.

Последний по времени появления и самый полный на сегодняшний день Частотный словарь (2009) опирается на коллекцию текстов Национального корпуса русского языка общим объемом 92 млн. словоупотреблений, представляющую русский язык последних шестидесяти лет. В корпусе собраны тексты различных функциональных стилей: художественная литература, публицистика, учебно-научная, официально-деловая, производственно-техническая и церковно-богословская литература, электронная коммуникация, реклама, бытовая литература, а также устная непубличная речь. Структура словарной статьи словаря предусматривает разные виды статистической информации, в том числе общую частоту употребления слова (количество употреблений на миллион слов корпуса) и ранг (номер слова в общем частотном списке всех слов словника), причем систему рангов имеют также словоформы, числа, буквы и знаки препинания.

2. Частотные словари фразеологии: terra incognita

В современной лексикографической практике отсутствуют частотные словари фразеологии. Это объясняется в первую очередь тем, что значительная часть идиом относительно низкочастотна. Кроме того, многие фразеологизмы характерны лишь для некоторых типов дискурса и речевых жанров. Например, идиома *город на Неве* является типичным журнализмом: в корпусе публицистики[1] она встречается 167 раз, а в корпусе русской прозы — всего 6. Наконец, употребление многих фразеологизмов определяется предпочтениями автора. Так, характерной особенностью идиостиля Достоевского является склонность к употреблению грамматического фразеологизма *по крайней мере* (685 вхождений по 35 художественным произведениям), при том что *по меньшей мере* у него практически не употребляется (три примера по 35 художественным произведениям).

Из сказанного следует, что для репрезентативности исходного корпуса требуется как достаточное количество текстов, так и их тематическое и жанровое разнообразие. Однако недостаточно механически увеличить объем корпуса и обеспечить большее тематическое разнообразие составляющих его текстов, что обычно рекомендуется для достижения репрезентативности в корпусных исследованиях. Существенно также учесть тот вклад, который вносит каждый конкретный тип дискурса в общий объем речевых произведений языка. Эта проблема пока не решена.

Наконец, имеются и объективные трудности формирования достаточно полного словника идиом, отражающего современный узус. Действительно, если представительный словник отдельных лексем можно получить, используя

[1] Здесь и далее речь идет о корпусах, созданных в отделе экспериментальной лексикографии ИРЯ РАН в середине 1990-х гг. Корпус публицистики 1990-х годов включает более 24 млн. словоупотреблений. Корпус русской прозы охватывает художественную литературу от 1960-х до 2000-х годов и включает более 33 млн. словоупотреблений.

имеющие словари и относительно небольшие по объему выборки текстов, то идиомы употребляются довольно редко. Имеющиеся фразеологические словари, во-первых, неполны и, во-вторых, существенно отстают от современного узуса. Извлечение идиом из текстов теоретически возможно, но практически неосуществимо — имеющиеся программы автоматической обработки выявляют устойчивые сочетания слов, однако результирующий список из-за большого объема и значительного информационного шума требует существенной редактуры и постоянного общения к контексту для установления того, использовано ли словосочетание идиоматично или нет.

3. Предпосылки создания частотного словаря идиом

В настоящее время в связи с развитием методов корпусной лингвистики, а также накоплением текстового материала появилась возможность решить многие из этих проблем. Во-первых, в результате многолетних исследований сформирован словник современных идиом, который может претендовать на полноту. Имеется в виду словник Словаря-тезауруса современной русской идиоматики (2007), который постоянно уточняется и дополняется. На данном этапе он насчитывает около 8000 русских идиом.

Во-вторых, в ходе работы над этим проектом были созданы три корпуса текстов, отражающие специфику различных сфер использования русского языка. В корпусе «Русская проза» представлены разнообразные речевые жанры — нарратив, бытовые диалоги, внутренняя речь (вплоть до потока сознания) и др. В корпусе публицистики собраны газетные и журнальные тексты второй половины 1990-х годов. Устная речь представлена опосредованно, в том виде, в котором она присутствует в корпусе «Русский детектив». Для более полного отражения этого типа дискурса в настоящее время создается корпус «Русская драматургия». Другие корпусы будут расширены за счет включения в них материала 2000-х годов.

4. Проблемы определения частотности идиом

Известно, что корпусы достаточно правдоподобно отражают частоту употребительных единиц языка, в то время как реальную частоту низкочастотных единиц отразить не удается. Поскольку идиоматика в целом относится скорее к низкочастотной части лексикона, возникает серьезная проблема. Она усугубляется тем, что разные говорящие в разной степени склонны к идиоматизации речи. Если один говорящий скажет *он пытался меня обмануть* или *он говорил мне неправду*, другой предпочтет иную форму: *он вешал мне лапшу на уши* или *он гнал пургу/лепил горбатого* и т. п. Идиоматичность (и соответственно, фразеологичность) речи — фактор индивидуальный, что хорошо видно по имеющимся источникам. Так, в Базе данных по современной идиоматике, включающей более 50 000 контекстов употребления идиом, произведения некоторых

писателей представлены большим количеством контекстов, а других — весьма незначительным. Например, из произведений В. Аксенова в базе имеются 603 контекста употребления идиом, Л. Петрушевской — тоже 603, Э. Тополя — 657, братьев Стругацких — 576, Саши Соколова — 542, Е. Гинзбург — 704, А. Солженицына — 652, В. Сорокина — 477, В. Войновича — 469, Ю. Трифонова — 471, Абрама Терца — 525, В. Приставкина — 473, В. Дудинцева — 435, С. Каледина — 343, Д. Быкова — 308, Б. Окуджавы — 310. С другой стороны, контекстов использования идиом в произведениях А. Марининой обнаружено 39, Л. Улицкой — 15, В. Шефнера — 5. Понятно, что в определенной степени это можно объяснить объемами представленных в базе произведений. Однако и при учете этого обстоятельства число употреблений идиом у «идиоматичных» и «неидиоматичных» авторов различается в десятки раз. Так, общий объем произведений В. Войновича в базе данных составляет в словоупотреблениях 428 400 токенов, а общий объем произведений А. Марининой, учтенных в базе данных, составляет 1 376 237 токенов. Тем самым, относительная частота употребления идиом в текстах В. Войновича составляет 0,001, а в текстах А. Марининой — 0,000028.

Таким образом, степень идиоматичности речи, проявляющаяся в частоте использования фразеологии, является индивидуальной чертой идиолекта каждого говорящего и пишущего. Данный факт необходимо учесть при создании корпуса для частотного фразеологического словаря. Действительно, если в нем будут широко представлены авторы, практически не употребляющие идиомы, то картина будет искажена, поскольку некорректно оценивать частоту какого-то явления у тех носителей языка, для речи которых это явление нетипично в принципе. Поэтому при подготовке корпусов художественной прозы, драматургии и детективной литературы следует исключить из них тексты тех авторов, которые не склонны к использованию фразеологии (для корпуса публицистики такой отбор по очевидным причинам невозможен). Порог частоты использования идиом для включения того или иного автора в корпус будет определен экспериментально.

5. Формирование словника

Основу словника словаря составят идиомы, зафиксированные в «Словаре-тезаурусе», за исключением идиом с обсценными компонентами. Это объясняется тем, что данные о частоте использования таких идиом были бы заведомо искажены: обсценная идиоматика является феноменом устной речи, ее отражение в литературных произведениях, даже включающих значительные фрагменты бытовых диалогов, крайне нерегулярно из-за табуированности этой группы слов русского языка.

Входом словарной статьи будет служить идиома в начальной форме. Начальная форма будет зависеть от типа идиомы — она может быть именной, адъективной, адвербиальной, глагольной и т. п. Будет использоваться та форма, которая представлена в «Словаре-тезаурусе…», поскольку для словника этого словаря уже была проделана соответствующая работа. Вопрос о том, будут ли варианты одной идиомы представлены в одной словарной

статье или в разных, пока не решен. В том случае, если они будут сведены в одну словарную статью, предстоит решить вопрос о критериях выбора заглавного варианта: это может быть наиболее частотный вариант, морфологически более простой вариант, наиболее полный по компонентному составу вариант и др.

Следует отметить, что, как и при создании частотных словарей лексики, этот этап работы включает разрешение омонимии. Ср. две омонимичных идиомы *на учёте*:

> **НА *УЧЁТЕ* I. (кто-л.)** Человек, мыслящий, думающий и т. п. ненормально или испытывающий затруднения в мыслительной деятельности и, как следствие, ведущий себя неадекватно из-за психического заболевания, которое документально фиксируется в соответствующем медицинском учреждении или из-за своей очевидности так осмысляется.
>
> *Мужчина, явно* ***на учете****, схватив ружье и, по предположениям, коробку дроби, закрылся в одной из комнат квартиры и категорически отказывался выходить.* [Корпус Публ.]
>
> **НА *УЧЁТЕ* II. (что-л. у кого-л.)** *нейтр.* Чего-л. так мало, что это требует постоянного контроля его наличия — в том числе письменной фиксации или так осмысляется.
>
> *Жанна — простая чертежница, у нее каждый рубль* ***на учете****.* [Корпус Публ.]

В тексте словаря полное толкование не предусмотрено, однако будет представлен краткий комментарий к значению, который позволит пользователю отождествить соответствующие употребления.

6. Типы информации в частотном словаре идиом

6.1. Частота по корпусам

В настоящее время рано говорить обо всех компонентах словаря и о деталях структуры словарной статьи. Укажем лишь те типы информации, которые заведомо будут отражены в словаре. Прежде всего, это относительная частота употребления идиомы во всех четырех корпусах, вместе взятых. Будет указана также относительная частота употребления в каждом из корпусов: Русская проза; Русская драматургия; Русская публицистика; Русский детектив.

Поскольку корпус «Русский детектив» и корпус «Русская драматургия» в значительной степени отражают особенности функционирования разговорной

речи, то будет отдельно фиксироваться и относительная частота употреблений в этих двух корпусах. Кроме того, целесообразно дать читателю относительную частоту употребления во всех корпусах, подвергшихся нормированию по степени «фразеологичности» (Русская проза, Русская драматургия и Русский детектив).

Как было уже отмечено выше, многие идиомы привязаны к конкретному речевому жанру. В то же время есть и идиомы, которые не ограничены жанровыми характеристиками. Так, идиомы *рано или поздно, ровным счетом, все до одного, без устали, между делом* нейтральны, то есть велика вероятность того, что они в равной мере употребительны в разных типах дискурса. Тем самым, встает задача упорядочить идиомы не только по их относительной частоте, но и по их распространенности в разных корпусах. Для этого в существующих частотных словарях используются разные статистические инструменты: в Чешском национальном корпусе это показатель Average Reduced Frequency (ARF) [Čermak, Křen 2005], в других частотных словарях — коэффициент А. Жуйана [Gries 2008]. Один из этих инструментов будет использован и в планируемом словаре.

6.2. Частотность компонентов идиом

Предполагается включить в словарь информацию о количественных характеристиках компонентов идиом, т. е. о том, сколько раз то или иное слово употребляется в составе разных идиом русского языка. Подобные данные имеют ценность как для исследования фразеологии, давая объективную информацию о лексическом составе идиом, так и для установления связей с лексикой, а также с морфологией и словообразованием.

Технология создания частотного словника компонентов идиом предусматривает следующие этапы. Все идиомы разбиваются на составляющие их компоненты, полученные словоформы сводятся к исходным формам лексем (леммам), которые снабжаются числовым индексом, показывающим частоту их употребления. В большинстве случаев правила сведения словоформ к исходной форме те же, что приняты в современных словарях (см., например, Частотный словарь русского языка (1977): 11–19). Так, для существительного это им. падеж ед. ч., для прилагательного — им. падеж ед. ч. мужского рода, для глагола — форма инфинитива. Но есть и некоторые отличия, связанные со спецификой фразеологического материала. Степени сравнения прилагательных и наречий, видовые пары глаголов, причастия и деепричастия, супплетивные формы (*человек/люди, ребенок/дети*) считаются разными леммами. Уникальные компоненты представлены в словнике в том виде, в каком они употреблены в составе идиом: *зги (не видно ни зги), скаку (на всем скаку)*. Предусматривается разделение омонимов, например, *да* как союз и как частица будут иметь разные входы и, соответственно, разную частотность. Полученный таким образом словник компонентов будет включать все слова, входящие в состав идиом — знаменательные слова (в том числе имена собственные), уникальные компоненты и заимствованные слова, а также служебные слова — предлоги, союзы, частицы и др.

6.3. Фразеологическая активность

Словник компонентов идиом будет представлен в двух вариантах — алфавитно-частотном и собственно частотном (по убыванию частотности). С помощью Указателя, содержащего отсылки от любого компонента к идиоме, пользователь сможет самостоятельно извлечь информацию об идиомах, в состав которых входит тот или иной компонент, а также о том, в каких словоформах он реализуется и насколько часто встречаются эти словоформы. Само присутствие в словнике того или иного компонента является признаком его «фразеологической активности», а частотный показатель позволяет судить о степени этой активности. Понятие «фразеологической активности» может рассматриваться как новый параметр исследования фразеологии.

Тем самым становится возможным выявление центра и периферии рассматриваемого множества единиц. Классификация слов, входящий в состав идиом, по различным основаниям может дать интересные результаты, характеризующие фразеологию русского языка как систему. Словник компонентов и их количественная характеристика могут служить, например, основой выделения различных семантических классов слов, входящих в состав идиом. Можно будет также исследовать «фразеологическую активность» отдельных грамматических и морфологических категорий (например, суффиксов и префиксов). Таким образом, частотный словник компонентов идиом может служить «мостиком», соединяющим фразеологию с другими аспектами языковой системы.

Полученные данные окажутся чрезвычайно полезными и при сопоставительном изучении разных языков. Так, сравнение словников и частотных характеристик может выявить как черты, сближающие фразеологические системы двух языков, так и важные содержательные различия[2].

7. Заключение

Создание частотного словаря идиом — новая теоретическая и прикладная задача. Ее решение позволит:

- определить относительную частоту употребления подавляющего большинства русских идиом в целом по корпусу и в разных типах текстов;
- выявить группы общеупотребительных, редких, стилистически связанных фразеологизмов, а также «авторских фреквенталий»;
- оценить распространенность в идиоматике омонимии;
- получить количественное представление о грамматических типах идиом;
- наглядно представить лексический компонентный состав русских идиом, а также степень фразеологической активности отдельных компонентов идиом и типов компонентов.

---

[2] См., например, реализацию подобного подхода при сравнительно-сопоставительном изучении фразеологии русского и немецкого языков в [Райхштейн 1980: 87–99].

Литература

1. *Баранов А. Н., Вознесенская М. М., Добровольский Д. О., Киселева К. Л., Козеренко А. Д.* Фразеологический объяснительный словарь русского языка. — Москва, Эксмо, 2009.
2. *Баранов А. Н., Добровольский Д. О.* Аспекты теории фразеологии. — Москва, Знак, 2008.
3. *Баранов А. Н., Добровольский Д. О., Киселева К. Л., Козеренко А. Д.* при участии М. М. Вознесенской и М. М. Коробовой, под редакцией А. Н. Баранова и Д. О. Добровольского. Словарь-тезаурус современной русской идиоматики. — Москва, Мир энциклопедий Аванта+, 2007.
4. *Белякова И. Ю., Оловянникова И. П., Ревзина О. Г.* (сост.). Словарь поэтического языка Марины Цветаевой. — Москва, Дом-музей Марины Цветаевой, 1996.
5. *Засорина Л. Н.* (ред.). Частотный словарь русского языка. — Москва, Русский язык, 1977.
6. *Лённгрен Л.* (ред.). Частотный словарь современного русского языка. — Uppsala : Acta Univ. Ups., Studia Slavica Upsaliensia 32, 1993.
7. *Ляшевская О. Н., Шаров С. А.* Частотный словарь современного русского языка (на материалах Национального корпуса русского языка) — Москва, Азбуковник, 2009.
8. *Поляков А. Е.* Электронный словарь языка писателя (на примере языка А. С. Грибоедова) // Труды Международного семинара Диалог-99 по компьютерной лингвистике и ее приложениям. — Таруса, 1999.
9. *Райхштейн А. Д.* Сопоставительный анализ немецкой и русской фразеологии. — Москва, Высшая школа, 1980.
10. *Степанова Е. М.* Частотный словарь общенаучной лексики. — Москва, Изд-во МГУ, 1970.
11. *Шайкевич А. Я., Андрющенко В. М., Ребецкая Н. А.* Статистический словарь языка Достоевского. — Москва, Языки славянских культур, 2003.
12. *Шайкевич А. Я., Андрющенко В. М., Ребецкая Н. А.* Статистический словарь языка русской газеты (1990-е годы)/ Рос. акад. наук. Ин-т русского языка им. В. В. Виноградова. Т. 1. — Москва, Языки славянских культур, 2008.
13. *Штейнфельдт Э.* Частотный словарь современного русского литературного языка. — Таллин, 1963.
14. *Čermak F., Křen M.* New generation corpus-based frequency dictionaries: the case of Czech // International journal of corpus linguistics, 10, 2005.
15. *Gries S. Th.* Dispersions and adjusted frequencies in corpora // International journal of corpus linguistics, 13, 4, 2008.
16. *Josselson H.* The Russian word count and frequency analysis of grammatical categories of Standard Literary Russian. — Detroit: Wayne University Press, 1953.

ПРОЛЕГОМЕНЫ К ПРОЕКТУ ГЕНЕРАЛЬНОГО ИНТЕРНЕТ-КОРПУСА РУССКОГО ЯЗЫКА (ГИКРЯ)

Беликов В. И. (vibelikov@gmail.com)
ИРЯ РАН, Москва, Россия

Селегей В. П. (Vladimir_S@abbyy.com)
ABBYY; РГГУ, Москва, Россия

Шаров С. А. (s.sharoff@leeds.ac.uk)
University of Leeds, UK; РГГУ, Москва, Россия

В докладе рассматривается задача создания Генерального Интернет-корпуса русского языка (ГИКРЯ). Анализируются технологические, структурные, функциональные, контентные недостатки современных корпусов русского языка и обосновывается необходимость и технологическая возможность создания нового корпуса на основе русскоязычного Интернета. Определяются принципы сбора, классификации, разметки, необходимые параметры и функциональность ГИКРЯ.

Ключевые слова: сегментно-статистический анализ Интернета, автоматическое создание корпусов, автоматическая классификация текстов, автоматическая разметка текстов

Введение

Современная лингвистика и лексикография являются в значительной степени корпусной: именно корпуса текстов являются сегодня важнейшим источником сведений о живом языке. Это делает особенно важным исследование вопросов, связанных с адекватностью методик корпусных исследований, правильностью устройства самих корпусов и эффективностью инструментов доступа к ним.

Авторы доклада являются сторонниками той точки зрения, что задача получения полноценного инструмента для корпусных лингвистических исследований все еще весьма далека от полноценного решения.

Доклад продолжает дискуссию, начатую на Диалоге-2011 статьей одного из авторов (Belikov 2011). В докладе делается попытка обобщить все известные проблемы с существующими средствами исследования текстов и анонсируется проект, основанный на идее создания Генерального Интернет-корпуса русского языка (ГИКРЯ), в котором эти проблемы могли бы быть преодолены.

Существующие корпуса относительно невелики по объему и их содержимое часто отражает случайные особенности их создания. Так, размеры русскоязычного Интернета на несколько порядков больше корпусов типа НКРЯ. Грубая

оценка на основе результатов Gulli, Signorini (2005) дает примерно 500 миллиардов слов русского языка индексированного поисковыми машинами в 2005. Это объясняет желание лингвистов проводить исследования на основе запросов к поисковым машинам. Но такие машины не предназначены для лингвистических исследований и частотная информация из них не всегда адекватна. В них также тяжело задавать лингвистически релевантные запросы.

Решением проблемы объем/функциональность явилось бы создание машины для лингвистической обработки русскоязычного Интернета, которая обеспечивала бы сбор достаточного количества страниц (сравнимого с индексом поисковиков), распознавание структуры текста на страницах, качественный морфосинтаксический анализ со снятием неоднозначности, автоматическую классификация текстов по темам и жанрам, а также по другим релевантным критериям метатекстовой разметки (география, возраст, пол автора, время создания). В этом интерфейсе будет возможно также визуализировать примерное положение текстов, интересующих нас, в сравнении с другими текстами Интернета, например, насколько далеко данные тексты отстоят от типичных новостей, либо от художественной литературы (Wilson, et al, 2011).

Содержательные и технологические проблемы существующих корпусов

Формально корпусом может быть названо любое собрание текстов с заявленным создателями принципами отбора, позволяющими оценить соответствие замысла и исполнения (так сказать, по законам, ими самими над собой установленными). Среди корпусов выделяются «универсальные», то есть такие, которые содержат языковой материал, адекватный по мысли создателей для практически любых исследовательских задач. Далее речь пойдет только об условно универсальных корпусах.

Несколько огрубляя, можно говорить о трех основных их типах с точки зрения состава и функционала:

1. Классические корпуса XX века, составленные вручную «закрытые» корпуса. Во многих случаях «корпус» рассматривается пользователями и даже авторами как материализация понятия «норма языка». К сожалению, далеко не всегда ясно, на чем основывается претензия на универсальность такого корпуса, и где ее «границы». Закрытые корпуса наиболее функционально развиты: исследователю может предоставляться возможность работать с подкорпусами, на которых произведено ручное снятие омонимии или проведена разметка, включая синтаксическую и семантическую. Такие подкорпуса позволяют искать и собирать статистику не только для отдельных слов, но и для выделенных при разметке лексических значений, синтаксических конструкций, целых семантических полей. Эти же размеченные подкорпуса используются для машинного обучения систем автоматического анализа текстов.

2. «Открытые», стихийно создаваемые без определенного плана массивы текстов, к каковым прежде всего относится интернет в целом. Несмотря

на отсутствие плана считается, что в интернете можно найти всё. Это и дает некоторые основания считать его особым универсальным корпусом. Что касается функционала, то он предоставляет весьма ограниченные и, увы, не слишком надежные средства доступа. Индексирование Интернета поисковиками, подсчеты частот, обработка запросов проводится на основаниях, далеких от лингвистических. Перифразируя известную цитату можно сказать «Сеть наша велика и обильна, а порядка в ней нет».

3. Некоторым компромиссом между этими подходами является бурно развивающаяся область исследований, которую можно назвать «Автоматическое получение корпусов из Интернета». Лингвистический статус автоматических корпусов, собранных обычно с помощью запросов по ключевым словам, не вполне ясен. В хороших случаях авторы подвергают их автоматической жанровой классификации, что дает некоторое представление об их структуре, а затем используют как универсальные ресурсы для лингвистических исследований (Sharoff, 2010; Bidhauer, 2012). Именно это направление и развивает предлагаемый проект.

Проблемы закрытых корпусов

Многие проблемы обсуждаются в уже упоминавшейся работе (Беликов, 2011), особенно в ее более полном электронном варианте[1]. Основной ее вывод: «It is a common belief that text corpora provide the best testing ground for solving any kind of linguistic problems. As far as grammar is concerned, this may be true, but if we focus on investigating the lexicon the results often appear to be rather superficial».

Определение «грамматика» в данном случае следует читать как «универсальная грамматика», поскольку в области лексикализованной грамматики мы сталкиваемся с теми же проблемами недостаточности материала.

Лингвистика и лексикография вынуждены обращаться к открытым ресурсам по двум основным причинам: а) из-за колоссальной динамики языковых изменений и б) из-за региональной, социальной и профессиональной «сегментации» языка, требующей дифференциального подхода к анализу языковых явлений и наличия нужного для получения надежных результатов объема языковых фактов.

Таким образом, основными содержательными недостатками закрытых корпусов являются «типологическая» неполнота или несбалансированность, часто случайность в отборе текстов и малый объем (особенно размеченной их части). Не принижая огромной роли замкнутых корпусов в решении многих задач, укажем на те, которые с их помощью решать не следует:

- Лексикосемантические и сильно лексикализованные синтаксические явления на языковой периферии (при этом, увы, с точки зрения надежности

[1] См. http://www.dialog-21.ru/digests/dialog2011/materials/ru/pdf/8.pdf

применения статистических методов верификации к периферии приходится относить очень многое).
- Исследование динамики языковых изменений.
- Социальная и региональная дифференциальная лексикография.
- Исследование терминологии.
- Сравнительные исследования текстов разных жанров.
- Фиксация и мониторинг языковой нормы.

Множественность нормы. Корпус и норма

До появления источников электронных текстов не было возможности объективно исследовать узус и выявлять на этом основании норму. Любая кодификация была субъективной. Однако, отождествление замкнутых универсальных корпусов (например, НКРЯ) с «нормой языка» требует весьма существенных оговорок.

Текстовое пространство сегментировано, и каждый сегмент, социаальный или географический, имеет свою норму.

Хорошим примером может служить различие профессиональной и общей нормы, ярко проявляющееся, например, в именовании общеизвестных биологических объектов. Так, данные, собранные в рамках проекта «Языки русских городов» показывают, что денотатом известной строки «Шумел камыш…» обычно признаются растения, известные биологам как рогоз и тростник. В то же время толковые словари чаще закрепляют именно профессиональную биологическую норму, которая практически всегда имела определенные (но в разных случаях разные) лингвогеографические обоснования.

Охота на корпусных снарков

В лингвистике и лексикографии как объекты описания, так и их свойства делятся де-факто на основные (как бы важные) и периферийные (как бы не очень важные). Важные являются относительно хорошо документированными, периферийные описаны не всегда верно, неполно и непоследовательно.

Замкнутые условно универсальные корпуса отличаются следующей особенностью: они хорошо подходят для описания основных объектов, и плохо — для периферийных.

Состав явлений, попадающих в корпусную «периферию», весьма неоднороден. Очень плохо, что по большому счету никто ее структуру не изучал. Речь не всегда идет об абсолютно редких объектах. Часто периферией оказываются объекты, быстро меняющие свое поведение (в том числе — появляющиеся и исчезающие), склонные к вариативности или очень неравномерно распределенные.

Не претендуя на создание термина, назовем для удобства ad hoc эти трудноуловимые объекты корпусными снарками[2]. Охота на снарков в замкнутом корпусе является по определению делом достаточно безнадежным.

---

[2] *They sought it with thimbles, they sought it with care,*
They pursued it with forks and hope,

Типичными представителми корпусных снарков являются:

- единицы региональная лексики;
- общая лексика за пределами частотного словаря в 30–40 тыс. слов.
- новые значения и новые модели управления;
- объекты актуального паремийного фонда;
- социолекты;
- распределение конкурирующих способов выражения (включая явления в грамматике).
- etc.

Так, лингвист, пытающийся на основании встреченного газетного примера «Мэрия согласовала нам маршрут движения…» провести корпусное исследование необычной модели с помощью НКРЯ, найдет лишь 2 примера разных лет. Никакого исследовательского вывода из неудачной охоты на снарка сделать, разумеется, нельзя.

Обращение к Интернету, даже без возможности задать грамматическое значение в запросе, немедленно показывает, что у «согласовать» быстро появляется новое значение ‘разрешить’ с похожей моделью управления.

Конкурирующие способы выражения это не только лексические синонимы (у которых реальное использование и стилистическая квалификация часто зависит от возрастных и социальных характеристик говорящего, ср. *захворать/заболеть/приболеть*). Фактическая (некорпусная) норма отражает «разноспрягаемость» синонимичных глаголов лазать/лазить: статистическое большинство предпочитает вариант *лазит*, но избегает формы *лажу*. Вариативность формообразования ряда глаголов хорошо известна, но здесь также проявлется неописанная «разноспрягаемость» (преобладает *плескаюсь в*, но *плещется в, плещутся в*).

Очевидно, что корпусные снарки являются одновременно и словарными: относительно них нет и не может быть и надежных словарных фиксаций.

В области хуже в целом описанной фразеологии соотношение идиом-конкурентов практически не исследовано. Так, словари отражают лишь *телячий восторг*, но в узусе даже старших возрастов он уже заметно уступает *щенячьему*.

Анализ языковой «периферии» вообще является делом чрезвычайно сложным. В замкнутых корпусах снарки легко путаются с сопоставимыми по частоте шумами и девиациями. При этом явления, плохо отслеживаемые корпусной статистикой, составляют в целом заметную часть текстов. Тут на место поиска «with forks and hope», примеров которого в лингвистических работах имеется достаточно много, должны прийти надежные корпусные методы, возможные только с использованием открытых корпусов.

They threatened its life with a railway share,
They charmed it with smiles and soap.
The Hunting of the Snark (An Agony in Eight Fits by Lewis Carroll)

Проблемы открытых ресурсов

Примеров корректно проведенных лингвистических исследований на открытых корпусах пока не так много. Назовем из последних лексикографические исследования на основании анализа блогов (Б. Л. Иомдин 2012 , В. И. Беликов 2012)

В некоторых случаях исследователям удается, опираясь на недокументированные возможности поисковых систем и/или во взаимодействии с их разработчиками, решать некоторые из указанных проблем, но систематического их решения ожидать вряд ли стоит.

Проблемы поиска стандартными поисковиками

Проблемы с использованием стандартных поисковиков широко известны, см. примеры в (Belikov, 2011), а также классический пример Veronis[3]:

Кратко эти проблемы таковы:

Надежность выдаваемой статистики

- Надежность выдаваемых цифр тем больше, чем цифры — меньше. А полностью доверять можно только тем цифрам, которые можно проверить полным просмотром выдачи.
- К числовым результатам не применимы аксиомы классической арифметики. Так, здесь часть вполне может быть больше целого (уточняя запрос получаем больше «результатов», чем для грубого запроса).
- Имеет место нестабильность: результат меняется во времени в произвольную сторону (не связанную с реальным изменением числа релевантных объектов) (Rayson, et al, 2012).

Очевидно, что информация о частоте запроса предназначена для совершенно других целей, чем лингвистические исследования. Результат зависит от множества разных факторов, не имеющих отношения к частоте употребления слов и конструкций (в настоящее время Яндекс стал говорить «Нашлось 42 тыс. ответов», а Гугл «Результатов: примерно 143 000 000» без указания единицы измерения).

Ненадежность в интерпретации результатов поиска хорошо известна, но лингвисты все равно пользуются ими, поскольку традиционные корпуса не дают возможности ответить на большинство запросов о частоте употребления даже среднеупотребимых конструкций (см. Крылов 2007). Отчасти помогают N-граммы (но такой базы на основе русскоязычного Интернета пока нет). Однако, в них, разумеется, не снята омонимия, что в отсутствии исходных контекстов ограничивает их применение как корпуса. Еще большая проблема

[3] http://blog.veronis.fr/2005/01/web-googlean-logic-en.html

таких списков — полная неясность с тем, как в них отражается дублетность, а это может исказить реальную картину на порядок и больше.

Снятие омонимии, синтаксическая разметка, язык запросов

Для поисковых машин не очень важна проблема снятия даже морфологической неоднозначности в самом запросе. Задавая вопросы поисковой системе, пользователи быстро приучаются вводить необходимые фильтрующие шум модификаторы. С другой стороны, сами системы поиска адаптируются к статистике запросов массовых пользователей. И вольно или невольно оказываются на поводу у ловких промоутеров и оптимизаторов, чем безусловно искажают реальность.

Совершенная иная ситуация возникает при использовании поисковиков в качестве устройств для лингвистического исследования Интернета (в частности, Рунета).

Здесь при исходном запросе '*помятые брюки*' полное равноправие вариантов «*не помяв новые брюки*» и «*подшивая брюки, помните, что сзади они должны доходить до каблука обуви*» вызывает серьезные проблемы для интерпретации статистики..

К сожалению, проблема в целом решается только параллельным снятием омонимии в запросе и в найденном тексте.

Для работы лингвистов корпус должен очевидным образом давать возможность задавать в запросе грамматические значения, пунктуацию, учет капитализации и прочее, к чему привыкли пользователи закрытых корпусов. Соответствующее направление развития на наш взгляд не совпадает со стратегией развития универсальных поисковиков, ориентированных в бизнес-процессе даже не столько на информационный, сколько на потребительский поиск товаров и услуг.

История с географией

Различие в подходах хорошо заметно на примере с географической атрибуцией интернет-страниц. В недавней работе (Volkov, Serdyukov, 2012) описывается алгоритм, применяемый в системе Яндекс.

Видно, что выбор параметров машинного обучения ориентирован на событийную, а не языковую специфику. В этом случае текст новосибирца о поездке в Москву будет атрибутирован скорее всего как московский. Сказать, правильно это или нет, не имея определения региональности и списка задач, ради которых она определяется, достаточно затруднительно[4]. Ясно, что для исследований регионального языка следует опираться на надежно установленные языковые региональные особенности, а не (только) на реалии, топонимы и проч.

Например, ФНС тщательно контролирует использование кассовых аппаратов. Запрос о *непробитых / невыбитых чеках* по делам, рассмотренным

[4] Другой вопрос, которого мы здесь не касаемся, это насколько последовательно реализовано то, что можно назвать информационно-региональной привязкой. Желающих адресуем к статье [В. И. Беликов. «Яндекс-регионы: найдется всё, выдастся без разбору»], являющейся дополнением к данной, в электронных материалах Диалога-2012.

в 17 Арбитражном апелляционном суде (обслуживает Пермский край, Свердловскую область и Удмуртию) даст не более четверти релевантных документов, поскольку в остальных речь будет идти о *неотбитых чеках*.

Проблемы дублирования и скрытого цитирования

Это одна из самых неприятных проблем, требующая колоссальных усилий от добросовестного исследователя. Тут есть вещи относительно очевидные — многократная перепечатка текстов, не очень очевидные — цитаты, и совсем не очевидные — паремии.

Исследователю, работающему с открытым корпусом, должны быть ясны приницпы, на которых поисковик относит тексты к дублетам. Проблему, какие дубли возможны и важны, а какие должны удаляться, мы полагаем совершенно нерешенной.

Неполнота и непоследовательность метатекстовой разметки

Печальным недостатком Интернета как открытого корпуса является отсутствие метатекстовой разметки. Даже дата создания является крайне ненадежным параметром, поскольку время публикации только для определенных сегментов Инернета совпадает с датой написания.

При этом жанровых однородных сегментов в Интернете очень мало. Имеющаяся в отдельных сегментах Интернета метатекстовая разметка непоследовательна. Скажем, даже в Классике библиотеки Мошкова беллетристика, поэзия и эпистолярий не разделены. В Журнальном Зале тексты упорядочены по изданиям, но жанровой классификации нет. Блоги наиболее насыщены метатекстовой информацией, но и здесь сложно говорить о полноте и последовательности ее приписывания. Так, в некоторых регионах блоги mail.ru не размечаются по возрасту, что делает практически невозможным проведение исследований по этом параметру. Кажется, что метатекстовая разметка в блоге имеет характер бонуса, а не обязательного параметра, за который отвечает разработчик поисковика.

Структурная неоднородность страниц

Страница Интернета в общем случае является сложно устроенным документом, с большим многообразием типов структур. Имеются два типа неоднородностей:

- Информацию на странице можно разделить на существенную для анализа постоянную «ингерентную» часть и динамическую составляющую: рекламу, новости, общую для нескольких страниц служебную информацию. К сожалению, страницы часто индексируются целиком.
- Часто структурно неоднородна и «ингерентная часть» страницы. Так, в Классике библиотеки Мошкова тексты авторов и комментарии к ним (обычно второй половины XX века) хранятся одним документом[5]. Очевидно, что языковые свойства этих разделов сильно различаются.

[5] На ФЭБе это разные документы, хотя все «пушкинское» лежит вместе и никаким сужением адреса эти части не разделить.

Программа проекта

Перечисленные выше проблемы можно решить только с помощью корпусов нового типа, сочетающих необходимую полноту исследовательского материала с наличием релевантой лигвистической разметки и основанной на ней надежной статистики.

Процитируем еще раз работу (Belikov, 2011): «WWW contains some relatively homogeneous arrays of texts formed independently of linguists, in some cases emerging quite spontaneously…Frequencies of the same lexical items differ greatly from one segment to another, and this statistics is very significant for sociolinguistics. The main problem in applying the method of segmental statistics is the lack of a suitable instrument for automatic data processing". То есть, возможность детального исследования многих типов узуса есть, но реализовать эту возможность без специального инструментария можно только «точечно», лишь в отношении конкретных частных языковых фактов, при этом такое исследование сегодня является весьма трудоемким.

До недавнего времени такая задача казалась неразрешимой, однако, мы полагаем, что сегодня имеются уже все основания для успешной реализации проекта, отвечающего нуждам лингвистов, и такой проект для русского языка (Генеральный интернет-корпус РЯ) должен быть обязательно запущен.

Если обобщить все претензии и пожелания к имеющимся ресурсам и инструментам доступа к ним, то получится следующая программа для проекта.

- Генеральный корпус должен быть настолько большим, чтобы быть релевантным для решения задач дифференциальной лингвистики и лексикографии (лексикографии жанровых, социальных и региональных различий). Поскольку речь не идет о задачах информационного поиска, объем корпуса не должен быть равным объему всего русскоязычного Интернета. Тем не менее, он должен быть на несколько порядков больше современных условно универсальных корпусов РЯ (10–100 миллиардов словоупотреблений).
- Этот корпус должен представлять все существенные социальные, жанровые, тематические сегменты Интернета и давать статистически достоверную картину относительного распространения текстов данного сегмента в сети.
- Генеральный корпус должен обновляться синхронно с обновлением Интернета (постоянно). Для целей обучения могут фиксироваться некоторые его версии.
- Интерфейс к корпусу должен обеспечивать поиск и подсчет частот с учетом любых параметров метатекстовой разметки.
- Поиск должен обеспечиваться технологиями статической (на уровне индекса) и динамической (на уровне обработки выдачи по покрывающему запросу) автоматической лингвистической разметки, позволяющей искать и статистически оценивать любые параметры языковых структур.
- Корпус должен предоставлять возможность использовать для динамической разметки альтернативные лингвистические модели.

Рассмотрим, насколько достижимы эти цели на современном уровне технологий, и что должно быть сделано в рамках проекта по созданию подобного корпуса.

Получение корпуса нужного объема и состава

Генеральный корпус должен быть открытым подмножеством Рунета, постоянно подпитывающимся новыми образцами. Получение сбалансированного корпуса нужного объема потребует многих итераций работы краулера на основании параллельно модифицируемого сегментного классификатора и сегментной карты Интернета.

Используемый нами термин *сегмент* применяется для указания на хорошо выделяемые и однородные в отношении некоторых параметров метатекстовой разметки подмножества текстов в Интернете. Сегменты могут быть физически компактными, например, электронные библиотеки или блоги, или виртуальными, распределенными по интернету. Разумеется, наиболее эффективна работа с компактными сегментами, которые систематически пополняются релевантными текстами.

Одним из наиболее ценных физических сегментов является блогосфера. Это, безусловно, наиболее динамичный, актуальный и размеченный массив текстов. Кроме того, только здесь и на форумах есть представительная информация о диалоговом общении.

Метатекстовая классификация и разметка

Частично классификация может вестись по материалам метаданных в отдельных ресурсах (пол, регион, возраст автора, иногда и его образовательный уровень в блогах, время создания/перевода текста в литературных коллекциях), но основу метатекстовой разметки в корпусе такого объема может составить только автоматическая классификация страниц.

В настоящее время достаточно хорошо себя зарекомендовали методы классификации по темам и жанрам, использующие самые простые и доступные признаки, такие как частота морфологических тегов или символьных N-грамм (Sharoff, 2010; Sharoff, et al, 2010).

В этом вопросе, однако, у авторов нет полного согласия. Необходимо проводить серьезные исследования с различными априорными классификаторами, выбор которых отнюдь не однозначен. В частности необходимы исследования по параметрам автоматической региональной привязки.

В интерфейсе также предполагается визуализация различий между исследуемыми группами текстов (см. Biber, 1989). Например, на основе частотности поверхностных признаков текста в сравнении с аннотированной нормой можно провести многомерное шкалирование (multidimensional scaling) либо анализ методом главных компонент (principal component analysis), чтобы определить позицию данного текста по шкале аргументативности или нарративности и сравнить его со средним значением для данного сегмента Интернета (Wilson, et al, 2011).

Автоматическая лингвистическая разметка

Полная ручная разметка Генерального корпуса заведомо невозможна, а частичная — бесмысленна, если не говорить о каких-то отдельных узких

сегментах Интернета, позволяющих сделать компактный, но достаточно представительный подкорпус.

Нужны новые инструменты, позволяющие производить автоматическую лингвистическую разметку (лемматизация, аннотирование частеречных признаков и синтаксический анализ) текстов, не связанную уже с ограничением на объем.

Было показано, что современные технологии морфологического и синтаксического анализа (Sharoff, Nivre, 2011) могут успешно работать на корпусе в 1–2 миллиарда словоформ.

В лингвистической разметке полезно различать условно физический и логический уровни. Генеральный корпус должен быть физически размечен признаками, с которыми готовы работать большинство исследователей. Что касается логического уровня, то он обеспечивается динамической постобработкой базового запроса на основании лингвистической модели разметки, которая может иметь альтернативы. Так, вряд ли можно говорить о едином синтаксическом представлении.

Системы семантической разметки еще сложнее унифицировать. Назовем для примера разметки на основе PropNet, систему разметки, основанную на модели Смысл-Текст (Apresjan et al. 2006), автоматическую семантическую разметку в системе Compreno (Selegey, 2012).

Очистка и структурирование страниц

Очистка страницы имеет нсколько уровней: от избавления от форматирующей информации до определения подструктур документа (см. выше про неоднородность).

Это очень важная тема, на которой мы не можем здесь останавливаться подробно. В целом, имеется определенная аналогия между задачами распознавания структуры документа в оптическом распознавании и для структурирования интернет-страниц. В нашем случае помогает доступная HTML-разметка.

Нужно иметь представления о формате основных коммуникативных форм: форумы, блоги, чаты.

Новые типы информации, представляемые корпусом

Лингвистическая и метатекстовая разметка позволяют решать задачи, которые сегодня еще не доступны исследователям, работающим с корпусами. Приведем только 2 примера:

- Переход в оценке от отдельных вхождений к числу страниц и далее — документов, был очень важен. Но все же наиболее надежная из всех видов статистики — авторская. Особенно эффективно инкорпорированность лексики в идиолекты можно оценивать в блогосфере. Легко выяснить, сколько блоггеров употребили за свою историю конкретное слово (словосочетание и т. п.). Для новых выражений при погодовых срезах отчетливо виден, например, переход от *конкретных* (*чуваков*) к *реальным* (*пацанам*), от *тусоваться* к *тусить*. Но в отсутствии разметки приходится «изворачиваться». Так, лингвист, интересующийся распределением

авторов между *лажу/лазию*, должен смотреть *не лажу/ не лазию*. (без отрицания на *лажу* получишь в основном *лажу*). И можно, наконец, выяснить, кого больше: тех, кто говорит *лингвистика* или тех, кто говорит *языкознание*.

- Возможности выявления фонетической специфики. Использование поэтических и песенных сайтов (*stihija.ru, stihi-rus.ru, rupoem.ru, pesni.net* и т. п.) позволяет по размеру делать наблюдения над ударением (можно доказать, например, широкую распространенность «запрещаемого» ударения пико́вый, — оно обычно для всех контекстов, кроме пиковой дамы, но и *пико́вая дама* встречается), по рифмам выявлять относительную встречаемость [чн] / [шн] в словах типа скучно, булочная, конечно. Много *булошных* и не ожидалось, но торжество *коне*[ч]*но* удивило. Некоторые ресурсы (stihija.ru, например) производят сугубо графоманское впечатление. Но для лингвистики это скорее хорошо: человек «фонетизирует» без оглядки на традицию.

Открытые вопросы для изучения

Переход к открытым исследовательским корпусам с автоматической разметкой порождает новые важные вопросы :

1. В какой степени корпуса, размеченные автоматически, пригодны для целей лингвистики и лексикографии, и как следует оценивать качество их разметки. Несомненно, что интегральная оценка качества более или менее адекватна для компьютерно-лингвистических приложений (для информационного поиска), но в случае исследовательских задач нельзя полагаться на средние данные (что-то вроде средней температуры по больнице) и нужно применять дифференцированные подходы к оценкам качества разметки с учетом различных языковых явлений.
2. Обоснование выбора языков разметки. Чем дальше мы продвигаемся от морфологии и частных задач разрешения лексической неоднозначности (основанных на каких-то канонических системах лексических значений типа WordNet), тем больше возникает проблем, связаных с различием подходов к описанию синтаксиса и семантики. Именно поэтому в современной корпусной лингвистике давно уже принят подход, различающий физическую разметку корпуса и его логические разметки. Важно, чтобы проект Генерального Интернет-корпуса предусматривал возможность добавления альтернативных разметок.
3. Большой размер универсального корпуса не во всех случаях является благом. Необходимо изучать вопросы определения выделения исследовательского подкорпуса — такого, на котором наиболее эффективно проводить конкретное исследовавние, включая и машинное обучение (Gascó, et al. 2012).

Заключение

Проект ГИКРЯ ни в коей мере не является альтернативой проектам создания образцовых замкнутых условно универсальных корпусов русского языка (например, НКРЯ) или специализированных корпусов, ориентированнх на исследование определенных явлений (референции, аффазии, etc). Важно понимать, для решения каких задач какой тип корпуса эффективнее.

Реализация проекта ГИКРЯ невозможна без широких обсуждений и участия заинтересованных исследователей. У него есть все шансы стать «рамочным» проектом — настоящим полигоном для апробации новых методов компьютерной лингвистики и лексикографии, прежде всего — методов автоматической классификации и разметки.

References

1. *Apresjan J., Boguslavsky I., Iomdin B., Iomdin L., Sannikov A., Sizov V.* A Syntactically and Semantically Tagged Corpus of Russian: State of the Art and Prospects. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, 2006. pp. 1378–1381.
2. *Belikov V.* 2011. What are sociolinguists and lexicographers lacking in a digitized world? Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2011". Moscow. pp. 60–67.
3. *Belikov V.* 2012. Segment-statistical approach to Internet as a corpus (on the example of blogosphere's analysis) Available at: http://www.abbyy.ru/science/seminars/archive/
4. *Bidhauer F., Schäfer R.* COW und texrex: Gigatoken Webkorpora and Tools zur Ad-hoc-Webkorpuserstellung. 34. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, Frankfurt am Main, 2012.
5. *Gascó, G., Rocha, M., Sanchis-Trilles, G., Andrés-Ferrer, J., Casacuberta, F,* 2012. Does more data always yield better translations? EACL2012, Avignon, France.
6. *Gulli, Signorini,* 2005. The Indexable Web is More than 11.5 billion pages. WWW05, Chiba, Japan.
7. *Iomdin B., Lopukhina A., Piperski A.* et al. Thesaurus of Russian everyday life terminology: new problems and new techniques. Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2012". Vol. 1. Moscow. pp. 213–226.
8. *Rayson P., Charles O., Auty I.,* Can Google count? Estimating search engine result consistency. Proceedings WAC'7, April 17, 2012, Lyon, France.
9. *Selegey V.* 2012. On automated semantic and syntactic annotation of texts for lexicographic purposes. Proceedings of the International Conference "Euralex 2012"
10. *Sharoff S.* In the garden and in the jungle: Comparing genres in the BNC and Internet. In Alexander Mehler, Serge Sharoff, and Marina Santini, editors, Genres on the Web: Computational Models and Empirical Studies, pages 149–166. Springer, Berlin/New York, 2010.

11. *Serge Sharoff, Zhili Wu, and Katja Markert.* The Web library of Babel: evaluating genre collections. In Proc. of the Seventh Language Resources and Evaluation Conference, LREC 2010, Malta, 2010.
12. *Sharoff S., Nivre J.,* 2011. The proper place of men and machines in language technology. Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2011". Moscow. pp. 591–604.
13. *Volkov A., Serdyukov P.*, Unified Classification Model for Geotagging Websites. Proc. of WWW 2012 April 16–20, 2012, Lyon
14. *Wilson, J., Hartley, A., Sharoff S., Stephenson, P.* (2011). Advanced corpus solutions for humanities researchers. Proceedings of the Pacific Asia Conference on Language, Information and Computation 2010.
15. *Krylov S. A.*, 2007. On the frequency of invective statements with universal reference and their negative correlates ("Yandex" as assistance to WCIOM: an attempt of quantitative analysis). Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2007". Moscow.

СЕГМЕНТАЦИЯ ТЕКСТА В ПРОЕКТЕ «ОТКРЫТЫЙ КОРПУС»

Бочаров В. В. (bocharov@opencorpora.org),
Алексеева С. В. (alexeeva@opencorpora.org),
Грановский Д. В. (granovsky@opencorpora.org),
Остапук Н. А. (nataxan90@gmail.com),
Степанова М. Е. (mariarusia@gmail.com),
Суриков А. В. (ksurent@opencorpora.org)

Проект «Открытый Корпус» OpenCorpora.org

Сегментация текста на слова и предложения является необходимым этапом автоматической обработки текста. Последующие этапы анализа существенным образом зависят от принятых соглашений о делении на слова и предложения. В проекте Открытый корпус вручную была проведена сегментация текста объёмом более 600 тыс. текстоформ. Полученные данные были использованы в качестве обучающей выборки для создания модуля автоматической сегментации.

Ключевые слова: корпус, сегментация, токенизация, графематика, машинное обучение

1. Введение

Под сегментацией текста мы понимаем деление текста на слова и предложения. Этот этап является первым при решении задач автоматической обработки текста, т. е. выполняется до запуска морфологического и последующих этапов анализа. Работа последующих этапов в существенной мере зависит от того, как поставлены границы структурных единиц текста на этапе сегментации. Во многих случаях принятие решения о постановке границы в том или ином месте кажется очевидным. В то же время встречаются сложные случаи, которые заставляют задуматься о том, должна ли быть поставлена граница в данном месте или нет. Это в равной степени касается как деления текста на слова, так и на предложения. В рамках проекта OpenCorpora была проведена работа по сегментации текстов общим объёмом более 600 тыс. словоупотреблений. На полученных данных была обучена математическая модель сегментации текста на слова. В этой статье мы опишем используемые данные, требования к программному модулю сегментации, принятые в проекте OpenCorpora правила сегментации текста на структурные единицы, причины, по которым они приняты, и полученные результаты.

1.1. О проекте OpenCorpora

Целью проекта OpenCorpora является создание размеченного вручную корпуса текстов на русском языке, полностью доступного для исследователей (см. [1] [2]). При этом основными пользователями корпуса мы видим специалистов, создающих программное обеспечение для автоматической обработки текста. Мы предполагаем два возможных сценария использования: обучение математических моделей на материалах корпуса и использование корпуса как тестовой коллекции для оценки качества систем анализа текста, основанных на правилах. Оба сценария требуют, чтобы разметка была единообразной. Т.е. если в корпусе имеются несколько содержательно одинаковых примеров, то они должны быть разобраны одинаковым образом. Требование единообразия реализуется при помощи стандарта разметки, которому следуют все специалисты, размечающие тексты. Если со временем становится ясно, что стандарт необходимо изменить, то вместе с изменением стандарта изменяются и все соответствующие этому изменению примеры.

1.2. О составе корпуса

В проект OpenCorpora включаются свободно распространяемые тексты, опубликованные на условиях, совместимых с лицензией CC-BY-SA, используемой в проекте. Таковыми являются материалы, опубликованные под этой же лицензией, находящиеся в общественном достоянии и не являющиеся предметом закона об авторском праве. С одной стороны, это существенным образом сужает круг материалов, которые можно использовать, с другой — является необходимым для того, чтобы корпус целиком был свободно доступен.

Актуальный состав корпуса постоянно изменяется в связи с добавлением новых текстов и представлен на странице http://opencorpora.org/?page=stats .

2. Требования к программному модулю сегментации текста на различные структурные единицы в проекте Opencorpora

Программный модуль сегментации текста запускается первым при автоматической обработке текста и передает свои данные на последующие этапа анализа. Последующие этапы анализа могут существенным образом различаться в зависимости от конечной цели. Это может быть результат цепочки модулей лингвистического анализа текста, включающей морфологию и синтаксис, или подсчёт частот слов для построения поискового образа документа, вектора признаков, необходимых для классификации или кластеризации. Таким образом, при проектировании модуля сегментации не следует полагаться на то, что какие-то другие модули (например, модуль морфологического анализа) будут запускаться перед модулем сегментации, и, следовательно, нельзя рассчитывать на доступность результатов работы этих модулей.

При разработке модуля деления на слова мы исходили из того, что при обработке текста модули запускаются последовательно, и результаты работы более поздних этапов анализа не доступны на предыдущих этапах. Сама по себе последовательная модель обработки не является единственно возможной, однако организация двунаправленного взаимодействия модулей усложнит интерфейсы их взаимодействия и сделает эти модули менее взаимозаменяемыми, что нежелательно.

Входными данными для модуля сегментации является последовательность знаков (букв кириллицы и других алфавитов, знаков пунктуации, цифр, пробельных символов и других, предусмотренных кодировкой знаков). Таким образом, никакая лингвистическая информация по умолчанию не предполагается доступной во время сегментации.

Актуальным способом представления текста в памяти компьютера является кодирование по стандарту Unicode. И хотя не все тексты на русском языке в настоящий момент представлены таким образом, мы считаем, что программный интерфейс современных систем анализа текста должен следовать стандарту Unicode, т. к. другие часто используемые способы кодирования (cp1251, koi8-r) могут быть без потерь перекодированы в Unicode. Обратное неверно. Более подробное обсуждение формы нормализации текста, закодированного в Unicode, мы оставим за рамками этой статьи, поскольку считаем этот вопрос в большей мере техническим, чем содержательным.

На выходе модуль сегментации группирует знаки в единицы более высокого уровня. В настоящей работе единицами, на которые происходит сегментация, являются токены, предложения и абзацы. При этом предложения состоят из токенов, а абзацы — из предложений. Полученные единицы станут в дальнейшем объектами последующих уровней анализа. Токены, являющиеся словами, получат морфологическую интерпретацию. Токены, не являющиеся словами, будут отмечены соответствующим классом сущностей: знаки препинания, числа, цифробуквенные комплексы (названия моделей, мероприятий, …), интернет-адреса. Предложения станут объектом синтаксического анализа и получат интерпретацию в виде синтаксического разбора. Мы считаем, что на более высоких уровнях анализа не должна возникать необходимость дополнительно делить токены на части. Расчленить на минимальные единицы, а после — сгруппировать, если будет такая необходимость, кажется более приемлемым решением, чем решать задачу сегментации на части в различных модулях системы.

Таким образом, вопрос о делении цепочки знаков на токены становится вопросом о том, станет ли данная цепочка в дальнейшем объектом морфологического, а после лексико-сематнического анализа (если не делить) или синтаксического анализа (если поделить).

3. Правила сегментации текста на токены

В проекте OpenCorpora выполняется сегментация силами волонтёров. Для того чтобы обеспечить единообразие при сегментации разными участниками,

были разработаны правила сегментации, регламентирующие постановку границ. Далее будут описаны эти правила и причины, по которым эти они были приняты. Как уже было сказано выше, правила не являются заданными раз и навсегда. Если принимается решение изменить правила, то все соответствующие этим изменениям примеры также переразмечаются.

1. Пробелы всегда разделяют токены. Таким образом, не может быть токенов, содержащих пробел. Все цепочки, включающие пробел (например, числа, в которых пробел служит разделителем разрядов), но обозначающие одну сущность, будут объединены на последующих этапах анализа.

2. Знаки препинания, как состоящие из одного знака (точка в конце предложения и после сокращённого слова, запятая, вопросительный и восклицательный знаки, …), так и из нескольких знаков (многоточие, несколько идущих подряд одинаковых знаков препинания) являются отдельными токенами. Интерпретация многоточия как отдельного знака препинания, а не как последовательности из трёх точек, кажется целесообразным с точки зрения последующего синтаксического анализа.

3. Числа, включающие точку или запятую в качестве разделителя целой и дробной частей, обыкновенные дроби, числа со степенью, интернет-адреса (URL, e-mail, …), химические формулы считаются целыми токенами. С точки зрения лингвистического анализа текста, анализ структуры этих цепочек знаков не несёт смысла, а значение имеет только класс выделенной цепочки.

4. Имена собственные, содержащие внутри знаки препинания (“Яндекс. Деньги”, “О.С.П.-студия”, “Санкт-Петербург”, “Нью-Йорк”, …) считаются целыми токенами. Эти имена могут быть написаны самыми разными способами и могут включать не только точки и дефисы, но и другие знаки (@, #, $, …). На последующих этапах анализа кажется целесообразным интерпретировать эти имена тем же способом, как интерпретируются все другие названия городов, компаний и брендов.

5. Обозначения единиц измерения (“км”, “$”, “%”, …), а также “+” и “-” перед числами отделяются от числа и становятся отдельными токенами, даже если между ними и числом нет пробелов. На последующих этапах анализа эта конструкция будет интерпретироваться одинаково (как число с указанием единицы измерения и знака) вне зависимости от того, были ли там пробелы.

6. Частицы, написанные через дефис с другим словом (“они-то”, “подойди-ка”, “он-таки”, “точно-с”, “он-де”, ...) отделяются. Таким образом, цепочка “они-то” поделится на три токена: “они”, “-”, “то”. Это нужно для того, чтобы частица попала в поле зрения синтаксиса как отдельная единица.

7. Повторяющиеся слова, написанные через дефис (“ха-ха-ха”, “много-много”, “красный-красный”, ...) разделяются аналогично предыдущему случаю, чтобы данное явление рассматривалось при последующем синтаксическом анализе.

8. Существительные, написанные через дефис и не подпадающие под предыдущие случаи, разделяются, если словарной статьи с таким заголовком нет ни в толковом, ни в орфографическом словаре (фактически проверка выполняется по коллекции Яндекс.Словарей). Это правило введено для того, чтобы

отделить синтаксическую конструкцию приложения от составных слов, написанных через дефис.

4. Правила сегментации текста на предложения

Деление текста на предложения чаще всего очевидно: границы предложений проставляются после знаков препинания, обозначающих конец предложения (точки, восклицательного знака, вопросительного знака) перед заглавной буквой или концом абзаца. Необходимость в более детальных правилах возникает в связи с прямой речью и цитированием.

Прямая речь отделяется от слов автора, даже когда слова автора не начинаются с заглавной буквы, а в конце прямой речи нет знака препинания, обозначающего конец предложения (вертикальная черта обозначает границу предложения):

В эту минуту кашлянул ребёнок, и я различил голос Горчи, он спрашивал: |
— Ты, малый, не спишь? |
— Нет, дедушка, | *— отвечал мальчик,* | *— мне бы с тобой поговорить.* |
— А, поговорить со мной? | *А о чём поговорить?* |
(А. К. Толстой, “Семья вурдалака”, 1839)

Таким образом, выполняется деление на минимальные единицы, которые после могут быть сгруппированы в единицы более высокого уровня, если это нужно. В дальнейшем предполагается помечать, какие из выделенных предложений являются прямой речью, а какие — словами автора.

При цитировании, оформленном при помощи кавычек, границы предложений расставляются по основному правилу, т. е. по знаку препинания, обозначающему конец предложения и началу следующего предложения с заглавной буквы. При этом возможна ситуация, когда открывающая кавычка будет в одной предложении, а закрывающая — в другом:

Накалив печь в деревянных банях, «туда входили нагими и там обливались водой. | *Потом брали розги (веник) и начинали себя бить, и до того секли, что едва выходили живыми.* | *Но потом, окатившись холодной водой, оживали».* |
(“Холодная ванна возвращает силы”, “Частный корреспондент”, 2008 г., http://www.chaskor.ru/article/holodnaya_vanna_vozvrashchaet_sily_14)

Такой способ деления на предложения сохраняет синтаксически целые предложения, отодвигая на второй план указание на авторство, отмеченное при помощи кавычек. Мы считаем, что для задач лингвистического анализа первое важнее второго.

Если цитируется название художественного произведения, включающее в себя несколько предложений, то границы внутри названия не ставятся:

Роман “Опасные связи. Или письма собранные в одном частном кружке лиц и опубликованные господином Ш.де Л.в назидание некоторым другим” увидел свет в 1782 году. |

В приведённом примере название романа могло бы быть однословным, и это не повлияло бы на синтаксическую структуру предложения, если рассматривать название как единое целое.

5. Автоматическая сегментация на токены

Первоначально модуль сегментации на токены (токенизатор) был разработан для того, чтобы упростить работу волонтёров по расстановке границ. На настоящий момент он выделен в отдельный модуль на языке Perl и опубликован в библиотеке модулей CPAN [6]. Модуль токенизации разработан с использованием машинного обучения, и он переобучается при добавлении новых текстов в корпус.

В качестве обучающего множества используются все уже токенизированные вручную предложения корпуса. Каждое предложение представлено в виде пары *(s, T)*, где *s* — исходная, нетокенизированная запись предложения (строка), а *T* — упорядоченное множество токенов, на которые оно разбито. При этом *s* всегда можно получить, "склеив" все элементы *T* по порядку, вставляя пробелы, если нужно. Следовательно, для каждой позиции j в s можно определить, стоит ли после j-го символа граница токенов.

Модель представляет собой вероятностный бинарный классификатор: для каждой позиции j вычисляется некоторое количество логических контекстных функций, ее характеризующих (см. ниже), и из результатов вычисления составляется вектор, например, <0, 0, 1, 1, 0, 1>, если функций 6. Накапливая информацию о парах (вектор, наличие границы), после обработки всего обучающего множества получаем для каждого вектора оценку вероятности встретить границу токенов в данных условиях.

Таким образом, применение модели для токенизации текста выглядит так: в каждой позиции текста по тем же правилам, что и при обучении, вычислить вектор двоичных признаков, найти в таблице соответствующую вероятность и решить, достаточно ли эта вероятность высока для постановки границы в данной позиции.

Сходные модели применяются для сегментации русского текста на предложения ([3, 4]), причем во втором случае предлагается более сложный (и, повидимому, более эффективный) метод, использующий деревья решений. Нет никаких сомнений, что этот вариант применим и в нашем случае; мы планируем исследовать его в будущем.

5.1. Контекстные функции

Большинство контекстных функций работают с минимальным, символьным контекстом — левым или правым. Сюда, например, относятся:

- "является ли символ слева буквой кириллицы",
- "является ли символ справа дефисом",
- "является ли символ справа закрывающей скобкой любого вида" и т. п.

Для работы таких функций достаточно информации, содержащейся во входном тексте.

Кроме того, некоторые функции принимают во внимание всю цепочку непробельных символов, внутри которой находятся — неважно, оказывается ли она в левом контексте, правом или обоих. Это имеет смысл, когда такая цепочка может оказаться объектом некоторого класса, который подразумевает действие "всегда разделять" или "никогда не разделять", например:

- адресом интернет-ресурса (http://domain.com/a/b/c/page.html) или электронной почты (mail@domain.com),
- словарным словом с дефисом (*по-английски*) и т. д.

Некоторым из таких функций требуются внешние источники информации — например, список всех словарных слов с дефисом или список исключений. Таким образом, в этом месте имеется неполное соответствие требованию не использовать лингвистическую информацию на этапе сегментации. С другой стороны — полного использования морфологического словаря в данном случае так же не происходит.

5.2. Порог отсечения

Принцип работы модели подразумевает, что каждой позиции нетокенизированного текста приписывается некоторая вероятность встретить в этой позиции границу токенов. Очевидно, что для разбиения этого недостаточно, так как необходимо привести каждую вероятность к бинарному значению (разбивать или не разбивать), для чего, в свою очередь, необходимо выбрать порог отсечения — т. е. значение вероятности, начиная с которого следует считать границу "достаточно" вероятной.

Стратегия выбора порога отсечения в целом представляется понятной: необходимо подобрать такое значение, при котором токенизатор будет показывать наибольшую точность работы (в широком смысле) при разбиении предложений обучающего множества (см. раздел "Оценка качества"). Это, очевидно, эффективная стратегия в случае, если в дальнейшем предполагается полностью автоматическая работа токенизатора. Однако если по каким-то причинам важнее, допустим, максимизировать полноту за счет точности (иными словами, лучше поставить несколько лишних границ, чем пропустить границу), то порог имеет смысл снизить примерно до 0,01.

5.3. Оценка качества

Точность бинарных классификаторов (каковым является токенизатор) общепринято оценивать путем подсчета количества ошибок, которые модель допускает при классификации объектов из некоторого корпуса, при этом ошибки бывают первого рода (ложное срабатывание) и второго рода (пропуск

события). В нашем случае ошибка первого рода — неверно поставленная граница токенов, а второго — пропуск правильной границы. Если рассматривать токенизацию как задачу поиска границ, то можно воспользоваться терминологией информационного поиска и, зная количество ошибок обоих типов, вычислить точность и полноту классификатора.

При оценке мы воспользовались известной техникой кросс-валидации (cross-validation, в русскоязычной литературе встречается также перевод “скользящий контроль”) [5], которая подразумевает, что некоторое размеченное множество примеров (в нашем случае — весь доступный корпус, около 600 тыс. токенов) разбивается на N подмножеств, после чего производится N измерений, в каждом из которых очередное подмножество выступает как тестовое, а остальные в совокупности — как обучающее. Затем результат усредняется. Мы выбрали N = 10: это значение достаточно велико, чтобы усреднение имело смысл, но ещё не столь велико для того, чтобы объём тестового множества стал слишком мал.

Поскольку, как мы уже упоминали, в проекте принято соглашение, что пробел всегда является разделителем токенов, довольно бессмысленно считать все верно расставленные границы до и после пробелов. Поэтому такие случаи были исключены из оценки.

На рис. 1 представлена зависимость точности (P), полноты (R) и F1-меры от порога отсечения (шаг по оси абсцисс — 0,05, кроме того, добавлены значения 0,01 и 0,99).

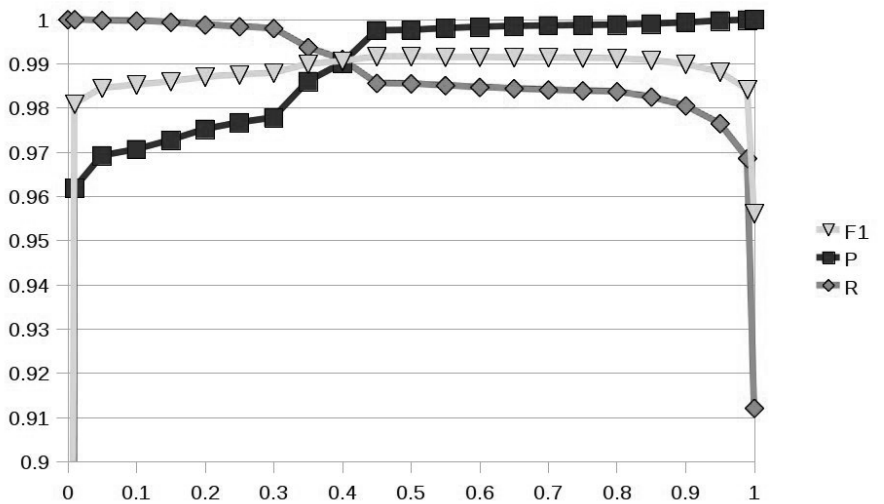


Рис. 1

Как и следовало ожидать, при увеличении порога отсечения точность монотонно растет, а полнота монотонно убывает. Интересно, что точность при пороге 0 составляет всего около 4%, а при пороге 0,01 уже более 96%. Полнота

при пороге 1 составляет около 91%. Максимум F1-меры достигается при пороге около 0,45–0,5 и составляет примерно 99,15%. Как показывает практика использования токенизатора в Открытом корпусе, это соответствует вполне удовлетворительному качеству токенизации, при котором большинство частотных случаев классифицируется верно. Самые спорные случаи классифицировать иначе, чем вручную, по-видимому, невозможно.

Заметим, что таким образом мы оцениваем, разумеется, не саму вероятностную модель, а конкретную ее реализацию, главным образом — набор контекстных функций, а также — в некоторой степени — качество обучающего множества.

Для сравнения мы проанализировали результат работы простейшего эвристического токенизатора с двумя правилами: всегда разбивать по пробелам, всегда разбивать на границе буквы и не буквы. F1-мера составила около 92%.

5.4. Преимущества и недостатки метода

Описанный метод, в принципе, обладает обычным преимуществом статистических методов: существенно проще — по сравнению с моделями, основанными на правилах — менять поведение токенизатора на некотором классе случаев (путем изменения набора контекстных функций).

При этом видятся три основных недостатка:

- невысокая скорость работы как при обучении, так и при классификации, не менее чем линейно зависящая от числа функций,
- зависимость от качественных (т. е. размеченных согласно единому стандарту) обучающих примеров,
- при собственном стандарте токенизации — необходимость изначально существующего размеченного множества примеров для первичного обучения.

Очевидно, что для некоторых задач этот метод токенизации не подойдет. Мы предполагаем два сценария, в которых его применение, скорее всего, оправдано:

- разметка объемного корпуса, особенно в условиях меняющегося по ходу стандарта токенизации,
- токенизация текста для задач, где детали стандарта токенизации не очень важны (подразумевается наличие уже обученной модели, см. ниже).

6. Заключение

В настоящей статье мы рассмотрели задачу сегментации текста, включая требования к её решению и место этого решения в системе автоматической обработки текста, правила сегментации текста при создании обучающей выборки, процедуру и результаты машинного обучения. Программный модуль,

реализующий описанную модель, доступен для свободного использования всем заинтересованным исследователям и разработчикам. Обучающая выборка, использованная для создания этого программного модуля, тоже опубликована, что даёт возможность разрабатывать иные способы решения указанной задачи.

Литература

1. *Bocharov V., Bichineva S., Granovsky D., Ostapuk N., Stepanova M.* Quality assurance tools in the OpenCorpora project // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25–29 мая 2011 г.). Вып. 10 (17). — М.: РГГУ, 2011.
2. *Бочаров В. В., Грановский Д. В.* Программное обеспечение для коллективной работы над морфологической разметкой корпуса // Труды международной конференции «Корпусная лингвистика — 2011». 27–29 июня 2011 г., Санкт-Петербург. — СПб.: С.-Петербургский гос. университет, Филологический факультет, 2011.
3. *Урюпина О.* Автоматическое разбиение текста на предложения для русского языка. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (4–8 июня 2008 г.). Вып. 7 (14). — М.: РГГУ, 2008.
4. *Кудинов А. С., Воропаев А. А., Калинин А. Л.* Высокоточный метод распознавания концов предложений. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25–29 мая 2011 г.). Вып. 10 (17). — М.: РГГУ, 2011.
5. *Kohavi R.* A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection // International Joint Conference on Artificial Intelligence, 1995.
6. *Суриков А.* Модуль Lingua::RU::OpenCorpora::Tokenizer. URL: http://search.cpan.org/~ksuri/Lingua-RU-OpenCorpora-Tokenizer/

ОПИСАНИЕ ГЭППИНГА В СИСТЕМЕ АВТОМАТИЧЕСКОГО ПЕРЕВОДА

Богданов А. В. (bidon@inbox.ru)

ABBYY, Москва, Россия

В работе рассматривается описание одного типа эллипсиса в системе русско-английского и англо-русского автоматического перевода. Приводится краткое определение данного типа эллипсиса (гэппинга), а также дается представление об архитектуре системы автоматического перевода. Рассматривается сам способ описания гэппинга, а затем подробно обсуждаются его плюсы и минусы. Приводятся примеры из реальных текстов, иллюстрирующие как плюсы, так и минусы описания.

Ключевые слова: эллипсис, гэппинг, машинный перевод, автоматический перевод

1. Рабочее определение гэппинга и теоретические вопросы

Гэппингом (англ. *gapping*) мы называем тип эллипсиса, при котором сокращается вершина одной или нескольких неначальных сочиненных составляющих. При этом могут подвергаться частичному сокращению и зависимые этой сокращенной вершины.

(1) *Петя любит Машу, а Маша — Петю.*

(2) *Традиционный салат «Оливье» будет стоять на столах 64% россиян, соленья — 54%, селедка под шубой — 46%.*

В (1) сокращается глагол (*любит*), являющийся вершиной правого конъюнкта сочинительной цепочки, при этом все его зависимые (*Маша, Петю*) остаются на месте. В (2) сокращается два глагола (*будут (будет) стоять*), являющиеся также вершинами неначальных конъюнктов. При этом подлежащее этих глаголов остается на месте, а другие их зависимые также подвергаются частичному эллипсису — на поверхности остаются лишь %, являющиеся прадочерними составляющими для сокращенных глаголов[1].

Сокращению не может подвергаться левый конъюнкт:

[1] В нашем понимании и в соответствии с приведенным нами определением, тип эллипсиса, обычно называемый left node raising (*Chomsky sent an offprint to Harris and a manuscript to his editor*), также является частным случаем гэппинга.

(3) ** Петя Машу, а Маша любит Петю.*

Гэппинг встречается во многих языках, и ему посвящено множество исследований, среди которых: [Ross 1970], [Jackendoff 1971], [Kuno 1976], [Neijt 1979], [Winkler 2005] и др.

Одна из самых сложных проблем при теоретическом описании гэппинга — тот факт, что сокращению может подвергаться материал, не образующий составляющую (как в (2)). Однако известны и строгие ограничения, накладывающиеся на гэппинг, — так, в [Ross 1970] впервые отмечается, что при гэппинге сокращению подлежит только такой материал, который образует непрерывную подчинительную структуру.

(4) а. *Маша хочет научиться курить сигареты, …*
б. *а Петя — научиться курить сигары.*
в. *а Петя — курить сигары.*
г. *а Петя — сигары.*
д. ** а Петя — научиться сигары.*

В (4б, в, г) сокращению подвергается непрерывная подчинительная структура, включающая правый конъюнкт — *хочет*; *хочет научиться* и *хочет научиться курить*. Тогда как в (4д) сокращению подвергается правый конъюнкт и его зависимое, не образующее с ним непрерывную подчинительную структуру — *хочет* <…> *курить*. И в этом случае сокращение оказывается невозможным.

2. Архитектура системы и проблемы описания гэппинга

Прежде чем говорить о способах описания гэппинга, кратко опишем архитектуру системы автоматического перевода, о которой пойдет речь.

Данная система перевода состоит из двух модулей — модуля анализа и модуля синтеза. На вход модуля анализа поступает текст на естественном языке, на выходе имеем семантическую структуру (на специальном семантическом языке); на вход модуля синтеза поступает семантическая структура, на выходе имеем текст на естественном языке.

Семантическая структура, упрощенно говоря, представляет собой дерево, в узлах которого находятся семантические концепты (обобщенные семантические единицы, не являющиеся объектами естественного языка), а связи помечены семантическими ролями.

(5) а.	*Мальчик*	*дал*	*девочке*	*яблоко.*
б.	[[$_{\text{Subject}}$мальчик]	дать	[$_{\text{Object_Dative}}$девочка]	[$_{\text{Object_Direct}}$яблоко]]
в.	[[$_{\text{Agent}}$BOY]	TO_GIVE	[$_{\text{Possessor}}$GIRL]	[$_{\text{Object}}$APPLE]]
г.	[[$_{\text{Subject}}$boy]	give	[$_{\text{Object_Indirect_to}}$girl]	[$_{\text{Object_Direct}}$apple]]
д.	*The boy gave an apple to a girl.*			

В (5а) представлен входной текст на русском языке. В (5б) — синтаксическая структура входного текста (промежуточный результат работы модуля анализа). Все связи в синтаксической структуре помечены типами (так называемыми поверхностными позициями). В (5в) представлена семантическая структура, в которой связи помечены семантическими ролями (концепты записаны заглавными латинскими буквами). В (5г) — синтаксическая структура выходного текста (с поверхностными позициями выходного языка), наконец, в (5д) — выходной текст на английском языке.

Очевидно, что для такой системы любой тип эллипсиса представляет некоторую сложность, так как все словоформы во входном тексте должны быть интерпретированы (им должны быть приписаны поверхностные позиции и должна найтись вершина, их присоединяющая). В простых случаях эллипсиса это решается восстановлением в месте сокращения некоторой условной вершины, к которой присоединяются имеющиеся во входном тексте зависимые.

Так, например, в случае регулярного эллипсиса *быть* в русском языке можно описать условия, при которых нужно всегда восстанавливать данный глагол.

(6) а. *Мальчик* *готов.*
б. [[$_{\text{Subject}}$мальчик] <быть> [$_{\text{Complement}}$готов]]
в. [[$_{\text{Object}}$BOY] TO_BE [$_{\text{State}}$READINESS]]

В (6) показан переход от сокращенной структуры входного текста (6а) к синтаксической структуре с восстановленным *быть* (6б) и далее к семантической структуре (6в). В упрощенном виде условия для восстановления *быть* в (6) выглядят так:

(6') […[Subject] …<быть>… [Complement]…];

Модуль анализа интерпретирует шаблон, записанный в (6') так: если во входном тексте имеется фрагмент, подходящий под этот шаблон, то нужно построить синтаксическую структуру с вершиной *быть* и данными зависимыми.

Простые случаи гэппинга вроде (1) можно описывать подобным образом. Например, можно использовать для этого такой шаблон:

(1') […[Subject]…Verb…[Object_Direct]…] & […[Subject]…<Verb>…
[Object_Direct]…]

При этом на месте сокращенного глагола можно восстанавливать лексему глагола, представленного в левом конъюнкте.

Однако шаблоны типа (1') неэффективны для описания гэппинга по следующим причинам:

- Зависимые сокращенного конъюнкта (как и левого конъюнкта) могут быть самых разных типов — понадобятся шаблоны для разных комбинаций

поверхностных позиций или какая-либо обобщенная запись поверхностных позиций.

- В большинстве случаев гэппинга имеется параллелизм поверхностных позиций зависимых левого конъюнкта и сокращенного конъюнкта — понадобится каким-то образом в шаблоне учитывать совпадение поверхностных позиций двух конъюнктов.
- Сокращению может подвергаться более глубокая структура, как в (2), и для таких случаев в шаблонах придется указывать все сокращенные элементы, которых может быть потенциально очень много.

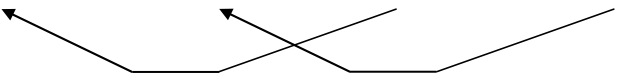
Приведенные причины делают практически невозможным написание обобщенного шаблона типа (1') для любых случаев гэппинга.

3. Описание гэппинга через контролирующие поверхностные позиции

Учитывая все вышесказанное, а также тот факт, что на гэппинг в русском и в английском языке накладываются схожие ограничения, была разработана другая система описания гэппинга — через контролирующие поверхностные позиции.

В упрощенном виде эта система выглядит следующим образом. Имеются две (в расширенном варианте — больше двух) технические поверхностные позиции с широким заполнением — то есть, такие, что в них могут попадать составляющие разных грамматических категорий (именные группы, глагольные группы и др.) и в разной грамматической форме (в разных падежах, с разными предлогами, союзами и т. п.). В этих позициях прописано свойство контроля. Контроль устанавливается между этими позициями и дочерними или прадочерними зависимыми левого конъюнкта. Покажем это на схеме.

(7) […[dependent]…Verb…[dependent]...] & [[Remnant1] <Verb> [Remnant2]]



В (7) показан шаблон с контролирующими поверхностными позициями. Ярлыком dependent обозначено множество поверхностных позиций для зависимых левого конъюнкта (будем записывать такие множества с маленькой буквы, в отличие от поверхностных позиций). Многоточиями обозначены другие возможные зависимые. Обозначения Remnant1 и Remnant2 — это технические поверхностные позиции для гэппинга, которые и являются контролирующими. Контроль от них идет в зависимые левого конъюнкта, он условно обозначен стрелками. Условием на установление контроля может быть согласование по определенному списку грамматических категорий (например, для существительных — по падежу и предлогу).

Покажем, как работает анализ и синтез примера (8а) по шаблону (7).

(8) а. *Мальчик рассказал девочке про синтаксис, а мне — про морфологию.*

б. [[$_{Subject}$ мальчик] рассказать [$_{Object_Dative}$ девочка; w] [$_{Object_Indirect_Про}$ синтаксис; x]] & [[$_{Remnant1}$ я; y] <Ellipted> [$_{Remnant2}$ морфология; z]];
y → w <падеж; предлог>
z → x <падеж; предлог>

в. [[$_{Agent}$ BOY] TO_TELL [$_{Addressee}$ GIRL; w] [$_{Theme}$ SYNTAX; x]] & [[$_{Remnant}$ I; y] <Ellipted> [$_{Remnant}$ MORPHOLOGY; z]];
y → w
z → x

г. [[$_{Subject}$ boy] tell [$_{Object_Dative}$ girl; w] [$_{Object_Indirect_About}$ syntax; x]] & [[$_{Remnant1}$ I; y] <Ellipted> [$_{Remnant2}$ morphology; z]];
y → w <падеж; предлог>
z → x <падеж; предлог>

д. *The boy told the girl about syntax and me — about morphology.*

В (8б) приведена синтаксическая структура, полученная в результате применения шаблона (7) к (8а). Латинскими строчными буквами помечены узлы структуры, участвующие в контроле. После скобочной записи перечислены две связи контроля — из поверхностной позиции Remnant1 в поверхностную позицию Object_Dative и из поверхностной позиции Remnant2 в поверхностную позицию Object_Indirect_Про. При установлении этих связей проверяется совпадение значений категорий падежа и предлога у контролера и мишени.

В (8в) приведена семантическая структура — результат работы модуля анализа. В ней сохраняются обе связи контроля, а зависимые сокращенного глагола получают технические семантические роли Remnant.

В (8г) приведена синтаксическая структура выходного языка. Обе связи контроля в ней сохраняются и реализуются через согласование по падежу и предлогу между контролером и мишенью. В результате получаем выходной текст (8д), в котором грамматическая форма *me* и *about morphology* — это результат работы согласования, то есть копирования значений соответствующих категорий с мишеней данных связей.

Такой способ описания гэппинга удобен тем, что позволяет в сложных случаях (как, например, в (2)) не восстанавливать промежуточные составляющие (между сокращенной вершиной и «остатками» гэппинга), но при этом синтезировать правильную грамматическую форму «остатков».

Продемонстрируем, как работает такое описание на более сложном примере с сокращенными промежуточными составляющими.

(9) а. *Информация синтетических счетов отражается в финансовой отчетности, а субсчетов — в приложениях к финансовой отчетности.*

б. [[$_{Subject}$ информация [$_{GenitivePostModifier}$ счет; w]] отражать [$_{Adjunct_Locative}$ отчетность; x]] & [[$_{Remnant1}$ субсчет; y] <Ellipted> [$_{Remnant2}$ приложение; z]]

y → w <падеж; предлог>
z → x <падеж; предлог>

в. [[$_{Object}$ INFORMATION [$_{Possessor_Metaphoric}$ BANK_ACCOUNT; w]] TO_MIRROR [$_{MetaphoricLocative}$ FINANCIAL_REPORTING; x]] & [[$_{Remnant}$ SUBACCOUNT; y] <Ellipted> [$_{Remnant}$ SUPPLEMENT; z]]
y → w
z → x

г. [[$_{Subject}$ information [$_{OfPostModifier}$ account; w]] reflect [$_{Adjunct_Locative}$ statement; x]] & [[$_{Remnant1}$ subaccount; y] <Ellipted> [$_{Remnant2}$ annex; z]]
y → w <падеж; предлог>
z → x <падеж; предлог>

д. *Information of control accounts is reflected in financial statements and of subaccounts — in annexes to financial statements.*

В (9б–в) показана работа модулей анализа и синтеза с теми же обозначениями, что и в предыдущих примерах. Некоторые составляющие из (9а) для краткости опущены. В отличие от (8), контроль в (9б) устанавливается не в дочернюю левого конъюнкта, а в прадочернюю (*счетов*), так как выполняется условие согласования именно с этой составляющей. Далее на синтезе в (9г) этот контроль обеспечивает согласование по падежу и предлогу между *subaccount* и *account*, а также между *annex* и *statement*. Благодаря чему получаем нужные грамматические формы *of subaccounts* и *in annexes*.

Для того, чтобы был возможен анализ типа (9б), в правилах контроля наряду с согласованием должны быть заданы возможные пути до мишени. Путем в данном случае мы называем цепочку поверхностных позиций последовательно зависящих друг от друга. Путь, обеспечивающий анализ (9б), выглядит так:

(10) Subject.GenitivePostModifier;

Путь, обеспечивающий синтез (9г), выглядит так:

(11) Subject.OfPostModifier;

Возможен гэппинг и с более глубоким сокращением, для которого могут понадобиться более длинные пути. Например:

(12) а. *Торговцам предоставили возможность заниматься* ***торговлей****, ремесленникам — своим ремеслом.*
б. Object_Direct.Clause_Infinitive_NoControl.Object_Indirect;

(13) а. *Rate of growth of credits to nonfinancial borrowers* ***in rubles*** *was of 101,9 %, in foreign currency of 103,3 %.*
б. Subject.OfPostModifier.OfPostModifier.Object_Indirect_in;

В (12) и (13) в пунктах а. представлены примеры из реальных текстов, а в пунктах б. — пути до мишени контроля, необходимые для описания гэппинга в этих примерах. Жирным выделены мишени контроля.

Очевидно, что в таких примерах восстановление всех сокращенных промежуточных составляющих было бы более трудоемкой и сложной задачей.

4. Возможные расширения предложенного описания

Описанную выше систему можно расширить несколькими способами.

Во-первых, число контролирующих позиций может меняться — встречаются случаи, когда параллелизм наблюдается только в одной паре поверхностных позиций, или наоборот — более чем в двух парах.

(14) *Объем привлеченных Сбербанком России депозитов увеличился* ***на 38,2%****, а в остальных кредитных организациях —* ***на 65,4%****.*

(15) *Внутренний кредит увеличился* ***на 544,6 млрд. руб.****, а за II квартал 2006 г. —* ***на 126,7 млрд. руб.***

В (14) и (15) показаны примеры, в которых только один «остаток» гэппинга параллелен одной из зависимых левого конъюнкта — *на 65,4%* и *на 126,7 млрд. руб.*. Вторая зависимая сокращенного глагола (*в остальных кредитных организациях* и *за II квартал 2006 г.*) в обоих примерах не имеет никакого коррелята в левой части. Точнее, в (14) таким коррелятом можно считать составляющую *Сбербанком России,* однако из-за отсутствия согласования по каким-либо категориям установить связь контроля между этими составляющими невозможно. Поэтому к примерам (14) и (15) не может быть применен шаблон типа (7). Однако подобные примеры можно описать с помощью шаблона с одной контролирующей позицией.

(16) […[dependent]…Verb…[dependent]…] & [[dependent] <Verb> [Remnant]]

В шаблоне (16) у сокращенного глагола (правого конъюнкта) указаны две зависимые — одна из них (dependent) может занимать любую поверхностную позицию из соответствующего списка (который можно дополнительно сузить), а вторая зависимая попадает в поверхностную позицию Remnant, из которой обязательно должен установиться контроль.

Другим возможным расширением предложенного описания может быть его применение к подчинительным конструкциям.

(17) *Молодые люди сегодня покидают дом быстрее, чем их коллеги поколение назад.*

(18) *Если в 1998 году у нас трудилось 1,1 тысяч иностранцев, то в прошлом году — 1,6.*

В (17) и (18) показаны примеры с эллипсисом, аналогичным гэппингу. Их отличие лишь в том, что сокращенный глагол не сочинен с главным, а подчинен ему. Такие случаи тоже могут быть описаны через контролирующие позиции с помощью шаблона вида:

(19) […[dependent]…Verb…[dependent]… [[Remnant1] <Verb> [Remnant]]]

После применения шаблона (19) к (17) должна построиться синтаксическая структура с двумя связями: (*их коллеги*) → (*Молодые люди*) и (*поколение назад*) → (*сегодня*). Аналогично для (18): (*в прошлом году*) → (*в 1998 году*) и (*1,6*) → (*1,1*).

Следует при этом принять во внимание тот факт, что возможные пути контроля, указанные в поверхностных позициях Remnant, в случае шаблона (7) идут от конъюнкта сокращенного глагола, а в случае шаблона (19) эти пути идут от родительской составляющей сокращенного глагола. Это различие можно указать, например, в самих шаблонах.

5. Недостатки предложенного описания

Предложенное описание гэппинга хорошо работает только в тех случаях, когда синтаксические структуры для данного примера на входном и выходном языке подобны, как во всех приведенных выше примерах. Если же синтаксическая структура входного языка при переводе должна каким-то образом трансформироваться, то предложенное описание может выдать плохой результат. Приведем примеры.

(20) *The boy was asleep. → Мальчик спал.*

Для перевода (20) в модуле анализа используется правило, которое производит следующее преобразование синтаксической структуры:

(21) [BE [$_{\text{Complement}}$ TO_SLEEP]] => [TO_SLEEP]

Теперь приведем пример с гэппингом, в котором *asleep* контролируется из позиции Remnant.

(22) а. *The boy was asleep and the girl — awake.*

б. [[$_{\text{Subject}}$ boy; w] be [$_{\text{Complement}}$ sleep; x]] & [[$_{\text{Remnant1}}$ girl; y] <Ellipted> [$_{\text{Remnant2}}$ awake; z]]
y → w
z → x

в. [[$_{\text{Experiencer}}$ BOY; w] TO_SLEEP; x] & [[$_{\text{Remnant}}$ GIRL; y] <Ellipted> [$_{\text{Remnant}}$ TO_STAY_AWAKE; z]]
y → w
z → x

В (22а) приведен пример, в котором сокращен правый конъюнкт, а одним из его оставшихся зависимых является *awake*, параллельное *asleep* в левом конъюнкте. В (22б) приведена синтаксическая структура, получившаяся в результате применения шаблона (7). Далее действует правило (21), заменяющее BE на TO_SLEEP и удаляющее составляющую в поверхностной позиции complement. После чего имеем семантическую структуру (22в), в которой связь контроля от TO_STAY_AWAKE идет к TO_SLEEP — то есть, к левому конъюнкту, а не к какому-либо из его зависимых. Такая семантическая структура не может синтезироваться стандартным образом, так как пути контроля в позициях Remnant не могут быть нулевыми.

Другим типичным примером недостатка предложенной системы описания гэппинга являются случаи, в которых между двумя «остатками» гэппинга должна быть какая-либо связь. Приведем пример.

(23) а. *The boy was stupid.*
б. [[$_{\text{Subject}}$ boy; x] be [$_{\text{Complement}}$ stupid; y]]
y → x

В (23) приведен пример, в котором между двумя зависимыми глагола *be* устанавливается дополнительная связь контроля. Она необходима в частности для того, чтобы согласовывать подлежащее и комплемент в таких конструкциях по роду и числу в русском языке (*Мальчик был глуп**ым**.*). Этот контроль прописан в поверхностной позиции Complement, то есть его установление обязательно для этой позиции.

Если подобные конструкции встречаются в контексте гэппинга, то возникает следующая проблема.

(24) а. *The boy was stupid and the girl — clever.*
б. [[$_{\text{Subject}}$ boy; w] be [$_{\text{Complement}}$ stupid; x]] & [[$_{\text{Remnant1}}$ girl; y] <Ellipted> [$_{\text{Remnant2}}$ clever; z]]
y → w (контроль прописан в позиции Remnant1)
z → x (контроль прописан в позиции Remnant2)
x → w (контроль прописан в позиции Complement)
в. [[$_{\text{Object}}$ BOY; w] BE [$_{\text{State}}$ STUPID; x]] & [[$_{\text{Remnant}}$ GIRL; y] <Ellipted> [$_{\text{Remnant}}$ CLEVER; z]]
y → w
z → x
x → w
г. [[$_{\text{Subject}}$ мальчик; w] быть [$_{\text{Complement}}$ глупый; x]] & [[$_{\text{Remnant1}}$ девочка; y] <Ellipted> [$_{\text{Remnant2}}$ умный; z]]
y → w (контроль прописан в позиции Remnant1) <падеж; предлог>
z → x (контроль прописан в позиции Remnant2) <падеж>
x → w (контроль прописан в позиции Complement) <род; число>
д. *Мальчик был глупым, а девочка умн**ым**.*

В (24б) приведена синтаксическая структура для (24а), полученная в результате применения шаблона (7). Перечислены все три связи контроля — две связи из контролирующих позиций гэппинга и третья связь — из позиции Complement в подлежащее. При переходе к семантической структуре (24в) все три связи сохраняются. Далее синтезируется синтаксическая стркутура (24г), в которой реализуется связь из позиции Remnant1 с согласованием по падежу и предлогу (в результате чего имеем форму *девочка*). Также реализуется связь из позиции Remnant2 с согласованием только по падежу (прилагательные при гэппинге не согласуются ни по каким другим категориям) и связь из позиции Complement с согласованием по роду и числу (в результате чего имеем форму *глупым*). При этом род и число для составляющей *умным* взять неоткуда. Фактически эти категории остаются неопределены, а в (24д) приведена форма с дефолтными значениями.

Другими словами, в конструкциях такого типа не хватает одной связи контроля — между *clever* и *girl*. Но ввести эту связь в структуру проблематично, так как она инициируется позицией Complement, а в данном случае *clever* подключается не в позицию Complement, а в позицию Remnant2.

Подобная проблема может возникнуть во всех случаях, в которых между двумя «остатками» гэппинга должна быть какая-либо дополнительная связь.

References

1. *Jackendoff, Ray* (1971) Gapping and related rules. Linguistic Inquiry 2, 21–35.
2. *Kuno, S.* (1976) Gapping: A functional analysis. Linguistic Inquiry 7, 300–318.
3. *Neijt, A. H.* (1979) Gapping. A contribution to sentence grammar, Foris:Dordrecht
4. *Ross, J.* (1970) Gapping and the order of constituents. In M. Bierwisch & K. Heidolph (eds.), Progress in linguistics: A collection of papers, pp. 249–259, The Hague: Mouton.
5. *Winkler, S.* (2005) Ellipsis and Focus in Generative Grammar. Mouton de Gruyter: Berlin/ New York.

О ПРОЕКТЕ СЛОВАРЯ ДИСКУРСИВНЫХ ЕДИНИЦ РУССКОЙ РЕЧИ (НА КОРПУСНОМ МАТЕРИАЛЕ)[1]

Богданова Н. В. (nvbogdanova_2005@mail.ru)

Филологический факультет Санкт-Петербургского государственного университета
Санкт-Петербург, Россия

В докладе представлены аргументы в пользу создания Словаря дискурсивных единиц современной русской речи. Под дискурсивными единицами в данном случае понимаются развернутые хезитативные конструкции, представляющие собой вербальное заполнение пауз колебания, столь свойственных устному речепорождению. Материалом для наблюдений и выводов служит Звуковой корпус русского языка (сбалансированная аннотированная текстотека и блок «Один речевой день»). Предлагаются первый вариант словника такого словаря и структура словарной статьи.

Ключевые слова: русская спонтанная речь, паузы хезитации, дискурсивные единицы, словарь дискурсивных единиц, хезитационная конструкция, Звуковой корпус русского языка

Описать язык — значит описать его словарь и грамматику. В этом смысле современный русский литературный язык описан достаточно полно и всесторонне. Перемещение внимания исследователей с языка на речь (в первую очередь устную), а также корпусный подход к сбору и систематизации живого речевого материала создают благоприятные условия для такого же описания речи как формы существования языка. Уже первые наблюдения над живым материалом показали, что проверка на корпусе самых, казалось бы, очевидных и общепринятых лингвистических утверждений иногда приводит к интересным и порой неожиданным результатам (см., например: *Богданова* 2010; *Богданова и др.* 2010, 2011), что позволяет поставить задачу пересмотра на корпусных данных *всех* накопленных к настоящему моменту лингвистических сведений. Это касается всех уровней языка — лексического (новые словари), морфологического и синтаксического (новые грамматики), фонетического, словообразовательного и пр. Не последнее место среди тех возможностей, которые открывает корпусный подход к анализу звучащей речи, занимает ее лексикографическое описание, как общее (см. о проекте *Словаря русской бытовой разговорной речи*:

[1] Исследование выполнено при поддержке гранта РФФИ «Изучение зависимости речевых характеристик от условий коммуникации (корпусное исследование на материале повседневной русской речи)» (проект 10-06-00300).

Богданова, Осьмак 2011), так и специализированное (см. там же о словарях *контекстных экспрессем, редуцированных форм* и *дискурсивных единиц*). Именно о необходимости и возможностях создания словаря дискурсивных единиц (ДЕ) современной русской речи и идет речь в настоящей статье.

Конкретный ***материал*** настоящего исследования был извлечен из двух источников:

1) *Национальный корпус русского языка* (НКРЯ) (основной и устный подкорпусы);
2) *Звуковой корпус русского языка* (ЗКРЯ) (два его блока: сбалансированная аннотированная текстотека — САТ — и «Один речевой день» — ОРД; см. о них подробнее: *Богданова и др.* 2008; *Степанова и др.* 2008).

Одной из ярких особенностей спонтанной речи, которая обдумывается и порождается в одно и то же время, т. е. в условиях временного дефицита, являются разного рода хезитационные элементы, в том числе и целые *хезитационные конструкции*, в которых вербализуются колебания говорящего в выборе речевых средств. Такие элементы, как показал анализ корпусного материала, весьма разнообразны — *(не) это самое*, *(я) не знаю*, *(я) (не) думаю (что)*, *в (на) самом деле*, *короче (говоря)*, *собственно (говоря)* и под. — и легко поддаются систематизации с точки зрения функций, выполняемых ими в спонтанной речи. Именно такие хезитационные конструкции и называются в данном случае *дискурсивными единицами* (в отличие от схожих терминов во многих других исследованиях — см., например: *Баранов и др.* 1993; *Дискурсивные слова…* 1998, 2003; *Кибрик А. А., Подлесская* 2005; *Шмелев* 2005; *Кибрик А. А., Савельева-Трофимова* 2008; см. также обзор соответствующих терминов в: *Дараган* 2000 и *Богданова* 2011).

В отличие от обычных *хезитативов* устной речи, как неречевых (*э-э, а-а, м-м*), так и лексических (*вот, ну, типа, как бы, это, как его* и под.), которые тоже относятся к ДЕ, нас интересуют в первую очередь вербализованные (более или менее развернутые) хезитационные конструкции, представленные в спонтанной речи двумя разновидностями.

Первая группа ДЕ восходит к лексически значимым фрагментам текста, и здесь можно, по-видимому, говорить о своеобразной их десемантизации:

- ***(я) не знаю*:**

  - *значит в этот день выходной / можно навестить / или какой-то вот […] родственников <u>там **я не знаю** /</u> по поводу ʃ событий может быть весёлых / и не очень весёлых* (САТ);

- ***(я) (не) думаю (что)*:**

  - о *ну / я не знаю / в шесть или в семь / **я думаю что** / я одну оставлю* (САТ);

- ***боюсь (что)*:**

- [Лена, жен / 52 / массажистка] *Только по другому поводу / да?* [Интервьюер, жен / ? / ?] *Да / **боюсь что*** [Лена, жен / 52 / массажистка] *Боже мой / а где они вот эти / которые и с головой и с руками?* [Беседа с массажисткой // Из материалов Санкт-Петербургского университета, 2006] (НКРЯ);

■ ***(…) скажем (…)*:**

- *этот мужчина сидит за столом / дабы вкусить то / что он поймал // **скажем так** чтобы в его ∫ <вздох> умосознании / осталось и-и восторг от пойманной рыбы / и от её вкусовых качеств* (САТ);

■ ***(ты) знаешь (что)*:**

- **В ну(:) / **ты знаешь** / если не лень / можно просто тупо с… считать / начиная с десяти //* (ОРД);

■ ***собственно говоря*:**

- ***собственно говоря** / выходной день ∫ э-э ∫ у меня нет выходного дня // в результате // вот и всё / что могу сказать / про выходной день* (САТ).

Возможно, к этой же группе можно отнести и такие единицы, как *понимаешь*, *короче (говоря)*, *так сказать*, *на (в) самом деле*, *как говорится* и некот. др. Все такие ДЕ часто выступают в тексте и в роли обычных вводных слов, что давно имеет лексикографическую фиксацию, однако в приведенных примерах это прежде всего хезитативы, что подкрепляется всей структурой текста (реплики), а часто и другими, сопутствующими, хезитационными маркерами (в примерах все они подчеркнуты).

Другая группа ДЕ восходит к незначимым фрагментам текста (словам-паразитам):

■ ***(не) (это) самое*:**

- *да да да да надо что-то … чтобы не очень претензионно было да и в то же время как бы … что-то надо нарядное # а оно **не это самое** / не долбанёт ?* (ОРД).

■ ***вот (…) вот*:**

- *привезли в Гагру / и там соответственно / до сих пор стоят эти ∫ эти гостиницы / и **вот этот вот** ∫ вот этот как раз ресторан* (САТ).

В обеих выделенных группах дискурсивных единиц хорошо видна их *полифункциональность*. Так, помимо заполнения паузы колебания, такие ДЕ часто

выступают в спонтанном тексте еще и как *дискурсивный* (чаще *стартовый* или *финальный*, реже — *направляющий, продвигающий монолог вперед*) *маркер*:

- ***знаешь / это(:) самое*** */ *Н сидят на месте / значит надо требовать / да / жалобы писать* (ОРД)*;*

- *университетскую // (…) она как-то называется / язык чего-то такое / (…) ну в общем короче / (м…м) *П каков наш / *В языковой день // **так скажем** / да ? *П вот мне дали такую штуковину / *П и я хожу / она на мне висит // *П все разговоры / *П записывает* (ОРД)*;*

- *и она просто / у нее там на на уровне подсознания срабатывает / нет / не хочу / потому что // я не знаю почему / **думаю что*** (ОРД).

Дискурсивные единицы могут выступать в спонтанной речи и как *маркер поиска* (глагола, имени, именного предиката, определения или обстоятельства; поиск при этом может быть удачным или неудачным; в случае удачного поиска найденное слово в примерах подчеркнуто):

- *испытываешь огромные-огромные / **вот эти вот** / <u>перегрузки</u>* (САТ)*;*

- *у нас сейчас () акция действует / скидка десять процентов на библиотеку // то есть это у вас получается как бы () **так скажем** <u>бесплатная сборка</u>* (ОРД)*;*

- *Настя / достань **это самое** и @ *Н @ выкини* (ОРД).

В случае с хезитационной конструкцией *вот (…) вот* можно говорить еще и о ее *дейктической функции*:

- *а потом / *П задумал **вот эту вот** задумку / построить **вот такой вот** склад для корабельных брёвен* (ОРД);

Часто хезитация в спонтанном тексте сопровождается еще и *метакоммуникацией*, когда с помощью дискурсивных единиц говорящий выражает сомнения в своих способностях правильно выразить ту или иную мысль, реализовать то или иное коммуникативное задание (напомним, что сбалансированная текстотека в составе ЗКРЯ включает спонтанные монологи, построенные в рамках серии коммуникативных сценариев — чтение, пересказ, описание изображения, рассказ на заданную тему, т. е. по определенному речевому заданию) или просто продолжить речь: здесь мы имеем дело с «текстом о тексте», или с «коммуникацией по поводу коммуникации»:

- *мне не нра… не нравится мне некие моменты / не нравится в какой-то степени даже очень больше для меня / вот я подумала*

*всё-таки / **я думаю** / **не знаю** / вот я бы наверное / если я такая хорошая как мне казалось он такой хороший / я б наверное так не поступила бы / ну вот // ну **не знаю** / во.()..во многих моментах // вот // ну знаешь как говорится нужно ценить то что имеешь* (САТ).

Нечасто, но встречаются в спонтанной речи и дискурсивные единицы в *функции самокоррекции* (элементы текста, подвергшиеся коррекции, подчеркнуты):

- *яркая солнечная погода // говорить можно? так был ярк… ∫ **это самое** ∫ был ∫ июльский день / вот / небо было чистым / безоблачным / солнце ∫ светило* (САТ).

В приведенных иллюстрациях хорошо видна полифункциональность дискурсивных единиц, при отчетливом преобладании все же функции заполнения паузы хезитации.

Именно наличие различных функций ДЕ в спонтанном тексте, часто совмещенных в одной единице, — при размытом, ослабленном или вовсе отсутствующем семантическом их наполнении — и вынуждает поставить задачу создания специального словаря таких единиц. Существующие лексикографические описания компонентов рассматриваемых хезитационных конструкций совсем не допускают возможности их употребления в выделенных функциях.

Так, в словарях русского литературного языка конструкция *это самое* в функции заполнителя паузы хезитации, естественно, не упоминается, и есть указание всего на одно ее разговорное употребление: «**САМЫЙ**, -ая, -ое, *мест определит*. 1. Употребляется при указательных местоимениях “тот”, “этот” для их уточнения в значении: именно, как раз. <…> || *Разг.* Употребляется в сочетании с местоимением “этот”, которое указывает на прежде названный или известный из предыдущего предмет, лицо. *Мне всегда хотелось сыграть в диккенсовском “Сверчке на печи”. Но я до сих пор не представляю себе, какие они **эти самые** сверчки*. Ю. Семенов. Майор “Вихрь”» (*Словарь русского языка* 1984: 25–26; подобные употребления в материал настоящего исследования, разумеется, не вошли).

Из всех выделенных в материале исследования функций слова *вот* лишь одна (функция указательной частицы) полностью соответствует словарным дефинициям, и то лишь в том случае, когда частица *вот* выступает в роли проклитики (второй компонент фонетического слова в примерах подчеркнут):

- *там значит в доме отдыха / где стоят такие курятники / где безо всяких удобств / **вот** эти м-маленькие э-э фанерные домики / там такое кафе шикарное / так дешево* (САТ).

Если говорить о структуре конкретной словарной статьи обсуждаемого словаря, то представляется, что она должна включать две обширных зоны: *семантическую*, сходную с толкованием компонентов ДЕ в обычных словарях

и дающую своего рода семантический фон для описания функционирования данной единицы в речи, и *функциональную*, в которой описываются все возможные функции данной ДЕ в естественной (по преимуществу устной) речи. В идеальном варианте такая статья должна содержать некоторые количественные соотношения выделенных функций, а также возможные корреляции употребления дискурсивных единиц в той ии иной функции с различными типами речи (*монолог — диалог/полилог*, *пересказ — описание — рассказ*) и/или характеристиками говорящих (*мужчины — женщины*, *старшие — младшие*, *экстраверты — интроверты* и под.).

Так, анализ функционирования слова *вот* в спонтанных монологах (блок САТ Звукового корпуса русского языка) показал, что самой частотной в материале исследования оказалась его *дискурсивная функция* (46,3%), особенно в таких ее разновидностях, как направляющий маркер (45,2% от общего количества употреблений *вот* в этой функции) и постпаузальный усилитель синтагматического членения (43,4%). Довольно частотна также *усилительная функция* (26,2%), особенно когда *вот* выступает в позиции проклитики (65,9% от общего количества употреблений *вот* в этой функции). Наименее употребительной в речи информантов оказалась *указательная (дейктическая) функция* слова *вот*, полностью соответствующая словарным дефинициям (1,4%) (см.: *Кислощук* 2011).

Анализ материала САТ, включающего конструкцию *(я) не знаю* (см. подробнее: *Хан* 2012), позволил установить, что наиболее распространенной в спонтанной речи оказалась эта конструкция в не свойственной кодифицированному языку функции заполнителя паузы хезитации (48,7%), особенно — в монологах-рассказах (47,4%). Существенно менее частотными оказались все другие, кодифицированные, употребления данной ДЕ:

- предикативная основа с придаточным изъяснительным (41,0%):
    - *купить какую-нибудь рубашку такую вот / **не знаю** фэнтэзи <u>или</u> фантази я так слово это и не прочитал* (САТ);
- компонент словосочетания с зависимой падежной формой (7,7%):
    - *это пейзаж / <u>названия</u> **не знаю** / а вот / дубы называется* (САТ);
- предикативная основа с зависимым союзным словом (2,6%)
    - *мы с женой / там пиво заказывали / и соответственно харчо брали / и харчо тоже ему очень нравилось / **не знаю** <u>почему</u> / вот* (САТ).

При этом существенной разницы в количестве употреблений исследуемой конструкции мужчинами и женщинами, а также экстравертами и интровертами выявить не удалось, хотя несколько замечаний в отношении гендера говорящих все же можно сделать. Например, как для мужчин, так и для женщин

пересказ сюжетного текста оказался наиболее простым коммуникативным заданием, поскольку количество того, чего информанты «не знают», здесь минимально по сравнению с описанием несюжетного текста. В речи мужчин исследуемая конструкция чаще употреблялась в описании несюжетного изображения (44,4%). В речи женщин нет явных «перепадов» от одного типа текста к другому, но максимальное количество исследуемых конструкций в итоге оказалось в свободных рассказах (33,3%).

Анализ функционирования в спонтанной речи хезитационной конструкции *это самое* (см. подробнее: *Богданова* 2011) выявил, помимо различных, уже рассмотренных выше, функций этой ДЕ, еще и две идиомы, включающих *это самое*, не требующих обязательного заполнения «пустой позиции» (как результата неудачного поиска слова) и свойственных почти исключительно разговорной речи: *а не пошли бы вы* ***в это самое****?* (НКРЯ) и ***как это самое*** (грамматические формы ДЕ при этом могут быть различны):

- *между прочим у французов тоже простая кухня / просто () это нам теперь здесь её () подают* ***как это самое*** */ как будто пре-парте какое-то* (ОРД и НКРЯ).

Оказалось также, что варианты данной дискурсивной единицы — *это самое / не это самое* — лишь частично являются антонимичными (положительная/ отрицательная):

- *Танечка ! будь ласка /* ***это самое*** */ спроси у Анечки* (ОРД)*;*

- *я говорю / свекровь поставила условие / если я … м-м* ***не это самое*** *… э-э не оформлю льготы на её квартиру / значит она нас по суду выписывает* (ОРД).

В значительной же степени отрицательная конструкция *не это самое* является независимой ДЕ, со своими функциями в спонтанной речи и своими особенностями употребления (подробнее о ней см.: *Богданова* 2012). Так, общим у обеих единиц являются их несомненные хезитационность и полифункциональность, с преобладанием функции поиска, в том числе неудачного, а также частое употребление в окружении пауз, как самостоятельной синтагмы:

- *нормально ? *П ничего там тебе /* ***не это самое*** *(…)* (ОРД).

- *вот сейчас они так будут нас футболить / вот он это … # вообще /* ***это самое*** */ *Н военнослужащий имеет право на получение … # да кто знал ? Зоя Станиславовна !* (ОРД).

Различающими же эти две ДЕ являются такие особенности конструкции *не это самое*:

1) нет функции *дискурсивного маркера* и *маркера самокоррекции*;

2) практически нет других грамматических форм данной конструкции (в материале нашелся лишь один подобный пример:

- *Почему ты* ***не этот самый****… как его?* (НКРЯ);

3) нет редуцированных форм ДЕ, таких, какие были выявлены для *это самое*:

- *много конечно есть примет // не знаю / хм / связанные с погодой // не знаю / что сказать / даже // ну как // то туман стеле… /* ***то самое*** *ну например по костру можно / определить погоду* (САТ);

- *а там дают очень какое-то блюдо большое с вкусной едой / если купите (…) вот… вот такую /* ***самый****… два с половиной литра пива* (ОРД);

4) очень редко *не это самое* выступает в роли чисто хезитационного маркера, подобные примеры нашлись только в письменной фиксации устной речи в Национальном корпусе русского языка:

- *Она говорит:* ***не это самое****, тебе не взять* [Календарь. Праздники. Часть 4 [Архангельская область, 1999] (НКРЯ).

Такая явная специфика конструкции *не это самое* требует для нее, по-видимому, отдельной словарной статьи в предполагаемом словаре.

Вся информация такого рода должна войти в словарную статью на конкретную дискурсивную единицу, что может придать обсуждаемому словарю весьма специфическую форму: это может стать своего рода собранием исследовательских *эссе* по каждой ДЕ. Более того, такая статья должна оставаться открытой для включения в нее новых данных или для коррекции имеющихся, поскольку Звуковой корпус русского языка, ставший основным источником материала для всех приведенных здесь наблюдений, продолжает пополняться, равно как продолжается и работа по его описанию.

Думается, что *Словарь дискурсивных единиц* может оказаться полезным не только специалистам-лингвистам, исследователям повседневной русской речи, но и в чисто практическом отношении: например, в практике перевода спонтанных текстов (хотя бы и в их письменной фиксации, в рамках художественного произведения, при передаче речи персонажей) на другие языки и в практике преподавания русского языка иноязычным учащимся, которые вынуждены учиться воспринимать и правильно понимать такую речь как устно, так и письменно, при чтении русскоязычных текстов.

Литература

1. *Баранов А. Н., Плунгян В. А., Рахилина Е. В.* Путеводитель по дискурсивным словам русского языка. М., 1993. 207 с.
2. *Богданова Н. В.* О корпусе текстов живой речи: новые поступления и первые результаты исследования // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 9 (16). По материалам международной конференции «Диалог» (2010) / Гл. ред. А. Е. Кибрик. М., 2010. С. 35–40.
3. *Богданова Н. В.* Конструкция (я) думаю (что) в русской спонтанной речи: соотношение разных функциональных типов // Проблемы социо- и психолингвистики. Выпуск 15. Пермская социопсихолингвистическая школа: идеи трех поколений. К 70-летию со дня рождения Аллы Соломоновны Штерн. Сборник статей / Отв. ред. Е. В. Ерофеева. Пермь, 2011. С. 266–275.
4. *Богданова Н. В.* Про не это самое (материалы к словарю дискурсивных единиц русской речи) // Слово. Словарь. Словесность: Литературный язык вчера и сегодня (к 300-летию со дня рождения М. В. Ломоносова). Материалы ежегодной Всероссийской научной конференции. Санкт-Петербург, РГПУ им. А. И. Герцена, 16–17 ноября 2011 года. СПб., 2012 (в печати).
5. *Богданова Н. В., Асиновский А. С., Маркасова Е. В., Степанова С. Б., Супрунова А. В., Шерстинова Т. Ю.* Звуковой корпус русского языка «Один речевой день»: пути пополнения и первые результаты исследования // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 9 (16). По материалам ежегодной международной конференции «Диалог» (2010) / Гл. ред. А. Е. Кибрик. М., 2010. С. 41–46.
6. *Богданова Н. В., Бродт И. С., Куканова В. В., Павлова О. В., Сапунова Е. М., Филиппова Н. С.* О «корпусе» живой речи: принципы формирования и возможности описания // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7 (14). По материалам ежегодной международной конференции «Диалог» (2008) / Гл. ред. А. Е. Кибрик. М., 2008. С. 57–61.
7. *Богданова Н. В., Осьмак Н. А.* О некоторых лексических «открытиях» на материале русской спонтанной речи (корпусное исследование) // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25–29 мая 2011 г.). Выпуск 10 (17) / Гл. ред. А. Е. Кибрик. М., 2011. С. 110–123.
8. *Богданова Н. В., Степанова С. Б., Шерстинова Т. Ю.* Звуковой корпус русского языка: новый подход к исследованию речи // Труды международной конференции «Корпусная лингвистика-2011». 27–29 июня 2011 г., Санкт-Петербург. СПб., 2011. С. 98–103.
9. *Дараган Ю. В.* Функции слов-«паразитов» в русской спонтанной речи // Труды международного семинара Диалог '2000 по компьютерной лингвистике и ее приложениям. Том 1. Теоретические проблемы. Протвино, 2000.
10. *Дискурсивные* слова русского языка: Опыт контекстно-семантического описания / Под ред. К. Л. Киселевой, Д. Пайара. М., 1998.
11. *Дискурсивные* слова русского языка: варьирование и семантическое единство. Сборник статей / Под ред. К. Киселевой, Д. Пайара. М., 2003.

12. *Кибрик А. А., Подлесская В. И.* Коррекция сбоев в устной спонтанной речи: Опыт корпусного исследования // Компьютерная лингвистика и информационные технологии: Труды международной конференции «Диалог'05» (Звенигород, 1–7 июля 2005 г.) М., 2005.
13. *Кибрик А. А., Савельева-Трофимова О. А.* Субклаузальные дискурсивные единицы в устной речи (на материале корпуса «Рассказы о сновидениях» // http://www.philol.msu.ru/~otipl/new/main/people/kibrik-aa/files/Subclausal@AR3-2008.pdf.
14. *Кислощук А. И.* Опыт систематизации функций слова вот в спонтанном монологе // Актуальные проблемы современной науки. Материалы международной научно-практической конференции научной сессии «XIII Невские чтения». СПб., 20–22 апреля 2011 г. / Под общ. ред. Н. И. Озеровой. СПб., 2011. С. 157–164.
15. *Словарь* русского языка в четырех томах. Том IV. С–Я / Гл. ред. второго издания А. П. Евгеньева. М., 1984.
16. *Степанова С. Б., Асиновский А. С., Богданова Н. В., Русакова М. В., Шерстинова Т. Ю.* Звуковой корпус русского языка повседневного общения «Один речевой день»: концепция и состояние формирования // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7 (14). По материалам ежегодной международной конференции «Диалог» (2008) / Гл. ред. А. Е. Кибрик. М., 2008. С. 488–494.
17. *Хан Н. А.* Конструкция (я) не знаю в русской спонтанной речи: соотношение разных функциональных типов // Мир русского слова, 2012 (в печати).
18. *Шмелев А. Д.* «Заполнители пауз» как коммуникативные маркеры // Язык. Личность. Текст. М., 2005.

АВТОМАТИЧЕСКИЙ МОРФОКЛАССИФИКАТОР РУССКИХ ИМЕННЫХ ГРУПП

Большаков И. А. (bolshakov34@mail.ru)
Независимый исследователь, Москва, Россия

Большакова Е. И. (eibolshakova@gmail.com)
Национальный исследовательский университет
Высшая школа экономики, Москва, Россия

Описывается морфологический классификатор русских именных групп с практически произвольным составом и длиной. Раздельно рассматриваются случаи одиночного существительного, группы не более чем с двумя изменяемыми словами и сверхдлинные группы. Классификатор испытан на тех 115 тысячах групп из 1–6 потенциально склоняемых слов, которые входят в словник большого электронного словаря русского языка КроссЛексика. Реализован также автономный вариант, опирающийся на встроенный. Он позволяет склонять именные группы, не входящие в КроссЛексику.

Ключевые слова: морфологическая классификация, русские именные группы, склонение групп, встроенный и автономный морфоклассификаторы

1. Введение

Практически любая система автоматической обработки текстов нуждается в морфологической классификации обрабатываемых слов. В русском языке для любых слов нужно знать их часть речи и флективный класс (= морфокласс), а для существительных — еще и род. Методы формального описания русской морфологии давно известны. Хороший обзор по теме содержится в работе [Sokirko], к которой мы и отсылаем.

Когда задумывался морфологический классификатор для большого словаря русских словосочетаний и смысловых связей КроссЛексика [Bolshakov], уже был внедрен в русскоязычную .версию Microsoft Word словарь А. Зализняка [Zaliznjak], а в двуязычных словарях МультиЛекс использовалась полиморфная модель А. Гельбуха [Gelbukh]. Однако задача классификации новых слов, не входящих в словари, еще не была решена. Не существовало и больших текстовых корпусов и массивов Интернета, для которых такая задача должна была неминуемо возникнуть.

Для КроссЛексики был нужен морфоклассификатор элементов словника, не известного заранее ни по составу, ни по объему. Классификатор должен

был осваивать новые морфологически не размеченные слова по мере их ввода в систему. Возникшая классификация предназначена для образования нужной флективной формы и полной морфологической парадигмы слова. При этом допускалось многолетнее совершенствование морфологических средств в процессе эволюции КроссЛексики.

Наша задача была сходна с задачей предсказания морфологии [Sokirko], но имела важнейшую особенность: объектом классификации могло быть как одиночное слово, так и целое словосочетание. В данной работе рассматриваются именные словосочетания (далее — именные группы, ИГ).

ИГ составляют 42% словника КроссЛексики и включают слова разных частей речи. В любом контексте ИГ играет синтаксическую роль существительного. ИГ может быть как одиночным существительным, так и словосочетанием из существительных, прилагательных, числительных и слов иных частей речи: ***точка*** *зрения,* ***вид*** *на жительство,* ***право*** *быть избранным,* ***свободная экономическая зона****, свободно* ***конвертируемая валюта****,* ***семь смертных грехов****,* ***теория*** *вероятностей и* ***математическая статистика***.

Внутри ИГ может быть несколько параллельно склоняемых слов (выше выделены шрифтом). Паттерны склонения для всех этих слов сводятся нами в единую таблицу, т. е. формально все они причисляются к существительным. Неизменяемость слова в ИГ понимается как сохранение вида словоформы при склонении группы в целом. Например, словоформа *речи* в группе *часть речи* является генитивом от *речь*, но эта форма сохраняется при любом вхождении объемлющей ИГ в контекст. Наличие нескольких изменяемых слов требовало синтаксического анализа элементов словника, а мы хотели его избежать. Облегчающим обстоятельством явилось то, что входной формой ИГ является ее именительный падеж.

Итак, на вход нашего морфологического классификатора подается ИГ в номинативе, а на выходе получается набор морфоклассов входящих в нее слов. Вся совокупность результатов классификации формирует внутренний морфословарь КроссЛексики. В рабочем режиме он позволяет восстанавливать любые косвенные падежи ИГ и выводить на экран всю морфопарадигму ИГ из шести падежей. На данный момент внутренний морфословарь включает 115 тыс. ИГ, наиболее употребимых в русском языке. Большинство многословных ИГ взято из Интернета.

Строя собственную морфологическую модель одиночного русского существительного, мы исходили из простейшей конкатенации <словоформа> = <псевдооснова> + <псевдоокончание>. Базовой же идеей предсказания морфокласса был принцип аналогии Г. Белоногова [Belonogov]: одинаковые цепочки последних букв (= финали), как правило, соответствуют одинаковому морфоклассу. Хотя финали длиной до пяти букв использовались широко, пришлось создать около сотни встроенных списков разной длины из слов, противоречащих аналогии. Для минимизации общего объема списков привлекалась статистика существительных разных морфоклассов.

В результате многолетней эволюции КроссЛексики, порядка 15% одиночных существительных ее словника оказались отсутствующими в словаре

Зализняка, но в рамках нашей морфосистемы склоняющимися правильно. Таким образом, нами построена морфосистема, альтернативная и эквивалентная системе словаря [Zaliznjak] (если не учитывать ударений и буквы «ё»). При этом наша система опирается на существенно обновившуюся за последние десятилетия лексику русского языка.

Из внутреннего морфословаря КроссЛексики и обслуживающих его подпрограмм был дополнительно сформирован автономный морфоклассификатор. Пользователь вручную вводит в него произвольную ИГ, а на экране получает ее правильную морфопарадигму. Кроме ИГ, содержащихся в КроссЛексике, автономный классификатор способен правдоподобно обрабатывать именные группы, отсутствующие в КроссЛексике, а также буквенные цепочки, отсутствующие в языке.

2. Классификация одиночных существительных

Простота приведенной выше формулы конкатенации обманчива, и морфоклассификатор русских существительных должен иметь многократно больше морфоклассов, чем указано в академических грамматиках.

- Большое количество существительных имеют беглую гласную. Поскольку беглых гласных в русском языке две, и порой они заменяются на *й* или ь, а следующих за ними разных согласных 13, для правильного морфологического описания нужны несколько десятков дополнительных классов.
- Существительные бывают одушевленные и неодушевленные. Их раздельное описание удваивает число морфоклассов.
- Существительные единственного и множественного числа мы рассматриваем раздельно, каждое со своей морфопарадигмой. Эта мера оправдана не только комбинаторными свойствами слов разных чисел, но и их различиями в наборах семантических связей. Число морфоклассов еще раз удваивается.
- Существительные, склоняющиеся по типу прилагательных, столь многочисленны, что покрывают все три грамматических рода, оба числа и практически все классы склонения прилагательных. Это прибавляет еще 46 классов.
- Именные группы могут формироваться числительными, у младших из которых одушевленный и неодушевленный варианты различны. Это еще десяток классов.
- Специфическим случаям типа *путь, дитя* или *Христос* приходится отводить отдельный класс. В итоге число морфоклассов достигло 230.

Дополнительно введены классы для существительных, входящих в многословные числовые группы типа *десять заповедей*. Здесь существительное *заповедей* имеет весьма необычную совокупность падежных окончаний {*ей, ей, ям, ей, ями, ях*}. Такие паттерны вынудили добавить еще

25 классов. Они эффективно обслуживают без привлечения сложных синтаксических правил множество ИГ, управляемых «малыми» (2–4, 22–24, 32–34...) и «большими» (5–20, 25–30, 35–40...) числительными в виде цифр либо слов.

Приведем фрагмент таблицы окончаний различных морфоклассов:

```
/*1*/{"","а","у","","ом","е","у","у"},   /*порт,бак,барыш*/
/*2*/{"","а","у","а","ом","е","а","е"}, /*автор,критик,малыш,подлец*/
/*3*/{"","а","у","","ем","е","у","у"},   /*абзац,вкладыш*/
/*4*/{"","а","у","а","ем","е","а","е"}, /*муж,гаденыш*/
/*5*/{"","а","у","а","ым","е","а","е"}, /*Иванов,Путин*/
/*6*/{"а","ы","е","у","ой","е","ы","е"},/*лапа,папа*/
/*7*/{"а","ы","е","у","ей","е","ы","е"},/*улица,певица*/
/*8*/{"а","и","е","у","ой","е","ы","е"},/*река,нога,межа,девушка*/
```

Слева идут номера морфоклассов, справа — примеры существительных. Число падежей увеличено до 8, поскольку у небольшого, но достаточно частотного ряда существительных мужского рода есть частичный (*выпить* ***чаю***) и / или местный падеж (*в* ***аэропорту***). В классах, где дополнительных падежей нет, их окончаниями взяты таковые родительного и предложного падежей соответственно. Особый вариант аккузатива в выражениях типа *пойти в программисты* считается нами номинативом с особым предлогом *в4* и с заглавным вопросом «*пойти* ***в какие люди***?» в модели управления.

С аналогиями среди существительных тоже не все гладко. Так в паре {*восток, росток*} слова с одинаковой финалью длины 5 склоняются поразному, но хотя бы имеют одинаковые род и число. Однако, например, в паре {*метель, мотель*} с одинаковой финалью длины 4 род различен. У омонимов нередко бывает разная одушевленность: *оператор1* ‘человек’ Vs. *оператор1* ‘операция’, а в паре *виски1* ‘части лица’ Vs. *виски2* ‘напиток’ различны и род и число.

Соображения аналогии пришлось дополнить оценочной статистикой крупных классов существительных. Упомянем лишь три учтенных нами наблюдения.

- Неодушевленных существительных больше, чем одушевленных в 2–6 раз, в зависимости от финали. Поэтому, если вводится новое слово и одушевленность его известна, мы помещаем его во встраиваемый список соответствующего класса, а по умолчанию классифицируем слово более правдоподобным способом как неодушевленное.
- Существительных множественного числа с финалью *ы*, склоняющихся по мужскому типу, заметно больше, чем склоняющихся по женскому типу. Поэтому, если появляется новое слово с женским типом склонения, оно помещается в списки раздельно одушевленных и неодушевленных слов, а по умолчанию новое слово на *ы* классифицируется по мужскому типу.
- Существительных женского типа с финалью <нешипящая согласная>+*ки* (*доски, пробки*...) больше, чем подобных мужских (*броски,*

ростки…). Поэтому в списки вносятся слова мужского типа, а по умолчанию новому слову присваивается женский.

В результат возникли порядка сотни встроенных списков, в основном, коротких. Наиболее длинный из них содержит одушевленные существительные единственного числа на твердую согласную (2250 слов). Встроенными списками суммарно покрываются примерно 12 тыс. слов, всего же в КроссЛексике сейчас около 48 тыс. одиночных существительных, так что нашим классификатором обеспечивается экономия примерно в четыре раза.

С учетом сказанного разработана процедура **SINGLE**(*noun, cls, num, gen*), где *noun* — входное существительное, например, *стена, стены, няни, нянечки, окно, вектор, строение, строения, учитель, учителя, лектор, студенты, студентки, бычки, шоссе, рагу, Горбачев, Михаил, Гарри, МГУ.* Остальные параметры процедуры выходные: *cls* — морфокласс существительного: 0,…,255*; num* — его число: единственное или множественное; *gen* — род: мужской, женский, средний или общий (только для множ. числа). Хотя в большинстве своем число и род однозначно определяются морфоклассом, это не так, например, для существительных на ***а*** типа *папа* или *бомбила* и для аббревиатур. Формируемые род и число ИГ нужны в разных частях КроссЛексики для проверки согласования.

Процедура **SINGLE** сначала разделяет существительные по их последним буквам. Далее существительные делятся исходя из предпоследней буквы. Далее используются более длинные финали и встроенные списки. Для примера, существительное *мама* получает правильный морфокласс вместе с сотнями других слов с финалью *ма* после предварительной проверки отсутствия в нем финали *дома* (как у слов *дома, автодома, детдома*…) и несовпадения со следующими словами: 1) *полкилограмма*, 2) *бельма*, 3) *письма*, 4) *Дюма*, 5) из короткого списка *далай-лама, Дима, Ерема, Кузьма, Обама*… 5) из короткого списка *закрома, корма, терема, тома*… После всех проверок с отрицательным результатом слову *мама* присваивается нужный класс по умолчанию.

3. Проверка и исправление морфокласса

Классификация одиночных существительных процедурой **SINGLE** не безупречна:

- Омонимам разной одушевленности (*конструктор, оператор, бычки*…) присваивается одинаковый неодушевленный класс, поскольку **SINGLE** не учитывает омонимии.
- Существительные, склоняемые как прилагательные, признаются одушевленными при наличии их во встроенном списке *обрученные, разнорабочие, трудящиеся*…, а по умолчанию считаются неодушевленными. В итоге, например, слово *военные* будет классифицировано как неодушевленное, что верно для группы *военные расходы*, но неверно при отдельном употреблении слова или в группе *демобилизованные военные*.

- Новые существительные множественного числа на *ы* по умолчанию относятся к мужскому неодушевленному типу (как *столы*), что не всегда правильно.
- Новые существительные с финалью <нешипящая согласная> + *ки* по умолчанию относятся к женскому неодушевленному типу (как *доски*), что тоже не всегда верно.

В то же время КроссЛексика включает обширные ресурсы, позволяющие проверить и даже исправить класс, присвоенный процедурой **SINGLE.** Это обширная база коллокаций (в целом их более 2 млн.), среди них типа *VN* (глагол → существительное: *нанять сотрудника*) и *NN* (существительное → существительное: *наем сотрудника*).

Вот как отражены в исходной текстовой базе КроссЛексики *VN*-коллокации существительного *оператор1*:

```
оператор1/«человек»                  передать2 ... ~у
  возлагать ... на1 ~а               посадить4 ~а
  заменить ~а                        пригласить ~а ...
  использовать ~а                    работать ~ом
  нанять ~а                          работать с2 ~ом
  обратиться1 к ~у                   стать ~ом
```

Цифры в конце слов означают номера омонимов, а тильда означает (псевдо) основу.

Для исправления неверного морфокласса используются противоречия окончаний существительного, записанных в базе коллокаций, и окончаний, соответствующих морфоклассу, определенному процедурой **SINGLE.** Противоречия возникают по двум причинам: (1) составителем базы коллокаций неверно введено окончание либо (2) **SINGLE** неверно определил класс. Чтобы автоматически различать эти случаи, используется булев показатель ***валидности***, дополнительно вырабатываемый процедурой **SINGLE** для каждого классифицируемого существительного: слово **валидно,** если в его морфоклассе нет сомнения в любых контекстах. Заметная часть валидных слов включено во встроенные списки.

При возникновении противоречий в случае валидного слова диагностируется ошибка составителя базы, и он должен исправить ошибку вручную.

В случае невалидного слова при отсутствии противоречий оно считается **верифицированным**. Если же противоречия есть, то программа сначала пытается исправить назначенный класс сама, без выдачи сообщений составителю базы коллокаций. Наиболее часто при этом неодушевленный класс исправляется на одушевленный.

Например, для слова *оператор1* при первоначально назначенном неодушевленном классе 1 обнаруживается противоречие в коллокации *возлагать на оператора* (здесь было бы тогда правильным *возлагать на оператор*). Противоречивы и коллокации *заменить / использовать / нанять оператора,*

но автоматически диагностировать это без предлога невозможно. В КроссЛексике имеется список из 63 соответствий {неодушевленный класс Vs. одушевленный класс}, где классу 1 противопоставлен одушевленный класс 2. Этот список и используется для коррекции: замена класса 1 на класс 2 устраняет для слова *оператор1* как обнаруженные, так и необнаруженные противоречия. Слова, у которых класс исправлен указанным способом, назовем **анимизированными**.

Существуют и другие исправимые противоречия. Так, если не включить существительное *лакуны* во встроенный список, можно исправить его «мужской» класс, полученный по умолчанию, на «женский», обнаружив пустое окончание в форме *лакун* (вместо **лакунов*). Слова, исправленные этим способом, назовем **феминизированными**.

Исправить класс слов *броски* или *ростки* с «женского» на «мужской» можно благодаря наличию для них в базе КроссЛексики псевдоокончаний *ов/ам/ами/ах* вместо *ок/кам/ками/ках*. Так исправленные слова назовем **маскулинизированными**.

Проверка окончаний для коллокаций типа *NN* не обеспечивает автоматических исправлений, но обнаруженные здесь противоречия тоже позволяют исправить базу коллокаций и / или встроенные списки**.**

После отладки встроенных списков и базы КроссЛексики получена следующая статистика валидации и верификации:

Валидные	57,7 %
Верифицированные	40,5 %
Анимизированные	1,6 %
Феминизированные	0,1 %
Маскулинизированные	0,1 %

Три последних позиции минимизированы переносом автоматически исправленных слов в соответствующие списки тогда, когда это допустимо. Но совсем исключить эти позиции принципиально невозможно из-за омонимов разной одушевленности и из-за возможной зависимости морфокласса от контекста.

Остается упомянуть еще омонимы, исправлять классы которых лучше подругому. В первую очередь это несклоняемые слова *ателье, бюро, интервью, казино, кафе, такси, фото, шоссе*... Условно мы приняли их единственное число первым омонимом, а множественное — вторым. **SINGLE**, не учитывающий омонимии, присваивает обоим омонимам единственное число, а затем второму омониму по номеру присваивается множественное число.

В итоге образуется файл BASICS из 20,6 тыс. одиночных классифицированных существительных, наиболее частотных в русском языке. Прочие процедуры КроссЛексики уже не классифицируют его составляющие заново. Пока не попавшие в BASICS 32 тыс. более редких одиночных существительных КроссЛексики постепенно переходят сюда по мере расширения базы коллокаций, а редкие одиночные пополняются новыми извне.

4. Классификация групп не более чем с двумя изменяемыми словами

Среди разнообразия ИГ в первую очередь выделим те, которые содержат не более двух изменяемых слов. Обнаружено четыре типа таких групп:

Дефисная группа. Разделителем частей ИГ является первый дефис, до которого нет пробела. Левая часть — изменяемое слово, правая часть — единственное слово, изменяемое или неизменяемое (*директор-распорядитель, шкаф-купе, страна-участница*). Сюда не относятся случаи слов с неразрывным дефисом типа *вице-президент*, *топ-менеджер* или *интернет-магазин.* Для различения этих случаев используются эвристики.

Сочиненная группа. Разделителем частей группы является сочинительный союз *и, или, да*. Обе части — одиночные слова, и любое может склоняться (*мама и папа, аукцион или конкурс, совет да любовь, готика и барокко*). Данный вариант покрывает и группы с сочинительным сокращением типа *аудио- и видеопродукция*.

Согласованная группа. Разделителем частей служит пробел. Слева и справа от пробела — одиночные слова, изменяемые или не изменяемые, которые согласованы по роду и числу (*конвертируемая валюта, конструкторское бюро, сочинительный союз, социальное государство, царь Петр*).

Левоуправляемая группа. Разделителем частей служит первый пробел. Левая часть — одиночное изменяемое слово, правая часть состоит из любого числа неизменяемых слов (*часть речи, вид на жительство, пресс-секретарь Белого дома*). Суммарное количество слов в таких группах может произвольно превышать установленный ранее предел 6: *Организация по безопасности и сотрудничеству в Европе.*

Именные группы, входящие в один из указанных типов, многочисленны: в КроссЛексике их 50 тыс. Технически оказалось затруднительным включить в указанные типы, например, группы *здоровый образ жизни, аутсорсинг и управление проектами, государства-участники СНГ, жилой фонд Москвы*, хотя в них ровно два изменяемых слова. Они отнесены к сверхдлинным группам.

Процедура **DOUBLE** служит для классификации ИГ описанных выше типов. Она получает на вход ИГ из произвольного числа слов и выдает один или два морфокласса, а также число и род, характеризующие группу в целом. **DOUBLE** последовательно проверяет соответствие ИГ одному из определенных выше типов. Когда соответствие найдено, к левой, правой или обеим частям группы применяется процедура **SINGLE**, что выявляет выходные морфоклассы, а также род и число группы в целом.

В качестве числа и рода дефисной группы берутся таковые левой части, например, *самолет-амфибия* — ед. муж., *ракета-носитель* — ед. жен., *общество-паразит* — ед. сред. В некоторых сочиненных группах, перечисленных списками, обе части имеют единый денотат, и тогда группе в целом присваивают его род, мужской (*друг и партнер, лжец и болтун*) или женский (*подруга и соперница*), а число берется единственным. Но по умолчанию денотаты различны, и группе присваивается множественное число.

В завершение **DOUBLE** проводит уточнение полученных параметров, бесконтекстное, диктуемое всей совокупностью хранимых в КроссЛексике многословных ИГ, либо контекстное. Так существительные *жучки / клещи / дипломаты / нищие / святые* всегда одушевляются, *банки / риски / белки* переводятся в склоняющиеся по мужскому типу, а *корма* и *вина* переводятся во множественное число. Наиболее частым приемом контекстного уточнения является анимизация левой части согласованной ИГ за счет правой части (*научные работники, лучшие люди*). В дефисных группах правая часть деанимизируется за счет левой (*город-герой*). Для многословных ИГ тоже приходится прибегать к встроенным спискам. Особо велик и пока еще растет список левых частей левоуправляемых групп (4500+ компонентов).

5. Классификация сверхдлинных групп

Среди многословных ИГ КроссЛексики встречаются и те, которые не охвачены описанными выше типами или содержат более двух изменяемых слов. Назовем такие ИГ сверхдлинными. Особо большое их количество оказывается среди десятков тысяч сочиненных пар, заимствованных из Интернета. При их морфологическом описании пришлось ограничить их состав. Конкретно, мы рассматриваем ИГ не более чем из шести слов, любое из которых может склоняться.

В целом они включают:

- ИГ с **тремя** изменяемыми словами (несколько тысяч): *Великая отечественная война, антибиотики и сульфаниламидные препараты, банковская гарантия и поручительство, логическое и творческое мышление*.
- ИГ с **четырьмя** изменяемыми словами (несколько сотен): *Великая октябрьская социалистическая революция, артезианская скважина и водонапорная башня, свободные экономические зоны и льготы*.
- ИГ с **пятью** изменяемыми словами (несколько десятков): *зеленые насаждения и малые архитектурные формы, десять заповедей и семь смертных грехов, единый социальный налог и страховые взносы, активная жизненная позиция и позитивный настрой*.

Статистически значимые примеры со всеми шестью изменяемыми словами типа *первый новый московский опытный моторный завод* в Интернете пока не встретились. В указанных ограничениях коллекция сверхдлинных ИГ в КроссЛексике открыта для пополнений. Однако внесение в словарь новой сверхдлинной ИГ требует ввода вручную в специальный текстовый файл EXLONG отдельной строки, содержащей как саму ИГ, так и некоторые ее показатели, позволяющие избежать синтаксического анализа. Вот примеры строк этого файла:

```
1 1 1 9 9 9 1 2 Великая отечественная война
1 0 1 1 9 9 2 4 антибиотики и сульфаниламидные препараты
```

```
1 1 0 1 9 9 2 4 банковская гарантия и поручительство
1 1 1 1 9 9 1 2 Великая октябрьская социалистическая революция
1 1 0 1 1 9 2 4 артезианская скважина и водонапорная башня
1 1 1 0 1 9 2 4 свободные экономические зоны и льготы
1 1 0 1 1 1 2 4 десять заповедей и семь смертных грехов
1 1 1 0 1 1 2 4 единый социальный налог и страховые взносы
1 0 1-0 9 9 2 4 кровати и шкафы-купе
```

Начальные *n* цифровых позиций, 3 <= *n* <= 6, кодируют изменяемость / неизменяемость (1/0 соответственно) каждого слова ИГ из *n* слов. Седьмая позиция кода занята показателем числа, восьмая — показателем рода ИГ в целом. Так, группа *Великая отечественная война* имеет единственное число и женский род, а большинство групп имеют множественное число и общий род. Заметим, что разделителем слов может быть не только пробел, но и дефис, смотри последнюю строку примеров. На данный момент файл EXLONG содержит 9840 строк.

На основе файла EXLONG процедура **MULTIPLE** обработки сверхдлинных ИГ каждому изменяемому слову в группе назначает его морфокласс, последовательно обращаясь к процедуре **SINGLE**. Назначенные номера классов запоминаются в файле EXLONG вместо показателей изменяемости «1». Далее совершаются автоматические бесконтекстные и контекстные правки, подобные указанным выше. Чаще всего — это анимизация слов, непосредственно предшествующих одушевленным словам.

Чтобы сформировать полную морфопарадигму сверхдлинной ИГ, из файла EXLONG извлекается ее последовательность классов склонения и по ним образуется вся форма со словами в одинаковом падеже. Пример полной морфопарадигмы дан ниже, где изменяемые окончания выделены:

имен *десят**ь** заповед**ей** и сем**ь** смертн**ых** грех**ов***
род *десят**и** заповед**ей** и сем**и** смертн**ых** грех**ов***
дат *десят**и** заповед**ям** и сем**и** смертн**ым** грех**ам***
вин *десят**ь** заповед**ей** и сем**ь** смертн**ых** грех**ов***
твор *десят**ью** заповед**ями** и сем**ью** смертн**ыми** грех**ами***
пред *десят**и** заповед**ях** и сем**и** смертн**ых** грех**ах***

6. Автономный морфоклассификатор

Автономный морфоклассификатор содержит в качестве базы внутренний морфословарь КроссЛексики. Процедура-движок сначала проверяет, имеется ли в нем входная группа. Если да, то группе присваиваются уже известные параметры.

Если ИГ в морфословаре отсутствует, то ее классификация выполняется процедурами **SINGLE** и **DOUBLE**. Правильно присваиваются классы огромному количеству пока не включенных в КроссЛексику групп произвольной длины.

Так, еще до включения в КроссЛексику, автономный морфоклассификатор просклонял правильно малоизвестные пока слова *диспорт* и *угги* и новое понятие *арабская весна*. Группа из 8 слов *перечень наркотических средств, психотропных веществ и их прекурсоров* в КроссЛексике отсутствует, но она тоже склоняется правильно.

Заодно правильно склоняются многие не существующие в языке группы, включая пресловутое словосочетание Щербы *глокая куздра*. Поэтому мы вправе заявить, что обрабатываемые группы имеют практически произвольный состав и длину. В литературе морфоклассификаторы многословных групп пока не известны.

Отметим неидеальную обработку автономным классификатором тех употребленных отдельно омонимов, у которых парадигма зависит от номера омонима. Внутри КроссЛексики правильно их классифицировать помогают именно эти номера. Требовать их знания от пользователя нельзя, и поэтому классификатор присваивает омониму наиболее правдоподобный класс. Полагая, что новые поступления будут удовлетворять приведенной выше статистике валидации и верификации, мы ожидаем правильный класс нового одиночного существительного в 98% случаев, а маловероятная ошибочная классификация проявится, скорее всего, только в винительном падеже.

7. Заключение

Предложен морфологический классификатор, сформированный в процессе пополнения объемлющей системы — большого электронного словаря словосочетаний КроссЛексика. Классификатор совершенствуется уже более 16 лет и в настоящее время практически стабилизировался в части алгоритма и встроенных списков. Он испытан на всех 115 тыс. именных групп, входящих сейчас в КроссЛексику. При этом для 46% групп достаточна процедура **SINGLE**, для 45% нужна **DOUBLE,** а для остальных 9% — **MULTIPLE**. Алгоритм и списки таковы, что ошибки классификации обнаружить уже не удается.

Автономный морфоклассификатор позволяет классифицировать с ничтожным процентом ошибок как включенные, так и не включенные пока в КроссЛексику именные группы и даже цепочки, отсутствующие в языке, но морфологически его имитирующие.

Наши классификаторы ориентированы на генерацию падежных форм. Анализа косвенных форм ИГ осуществим только совместно с другими частями речи. Пока в КроссЛексике анализируются только любые одиночные слова.

Выражаем признательность А. Ф. Гельбуху за встраивание морфологических процедур в КроссЛексику.

References

1. *Belonogov G. G.* On the use of a similarity method for automatic processing of textual information [Ob ispol'zovanii metoda analogii pri avtomaticheskoj obrabotke tekstovoj informatsii] Problems of cybernetics [Problemy kibernetiki]. Issue 28. Moscow: Nauka Publ., 1974.
2. *Bolshakov I. A.* CrossLexica: A large electronic dictionary of collocations and semantic links between Russian words [KrossLeksika — bolshoj èlektronnyj slovar' sochetanij i smyslovykh svjazei russkikh slov]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2009" [Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2009"]. Moscow, 2009, pp. 45–50.
3. *Gelbukh A. F.* Effectively realizable morphologic model of inflective language [Èffectivno realizuemaja model' morfologii flektivnogo jazyka]. Nauchno-tekhnicheskaja Informatsija [Scientific and Technical Information], series 2, #1, 1992, pp. 24–31.
4. *Sokirko A. V.* Bystroslovar': morphological prediction of new Russian words using very large corpora [Bystroslovar': predskazanie morfologii russkikh slov s ispol'zovaniem bolshikh lingvisticheskikh resursov]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" [Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2010"]. Moscow, 2010, pp. 450–45
5. *Zalizniak A. A.* Grammatical dictionary of Russian: Inflection [Grammaticheskij slovar' russkogo jazyka: Slovoizmenenie]. Moscow: Russkij Jazyk Publ., 1977.

РОЛЬ ДИСКУРСИВНЫХ СЛОВ В УПРАВЛЕНИИ ПОНИМАНИЕМ ТЕКСТА

Борисова Е. Г. (egbor@mail.ru),
Московский городской педагогический университет,
Москва, Россия

В статье рассматриваются служебные слова (в первую очередь, усилительные частицы и наречия), которые могут использоваться говорящим для управления пониманием текста, а именно для уточнения или отмены выводов и пресуппозиций слушающего. Показаны возможные действия говорящего, употребляющего эти единицы, в различных дискурсивных ситуациях. Из этого следует необходимость учета в семантическом описании не только эксплицитно выраженного смысла, но и наиболее вероятных пресуппозиций и импликатур.

Ключевые слова: семантическое представление, импликатуры, частицы, вводные слова, интерактивный подход

1. Служебные слова во взаимодействии участников коммуникации

При обращении к описанию речевой деятельности: прагмалингвистике, дискурсивному анализу — лингвисты обратили внимание на «словечки», т. е. служебные слова, которые ранее получали весьма неопределенное описание, сделанное по лексикографическому шаблону, выработанному для знаменательной лексики (глаголов, существительных и т. п.). Было сделано предположение, что большинство служебных слов: усилительные частицы, междометия, наречия, используемые в качестве вводных слов, союзы — используются для отражения информации, характеризующей различные аспекты взаимодействия между участниками общения. Так, Р. Ратмайр предложила термин «прагмалексемы», в число которых она включила русские усилительные частицы и немецкие Modalpartikeln (модальные частицы) (Rathmayr 1975). В дальнейшем описание названной лексики связывалось с ее поведением в дискурсе, что привело к появлению понятия «дискурсивные слова».

В рамках такого подхода получили новое описание многие служебные слова. Объем полученной информации позволяет теперь использовать дискурсивные маркеры для моделирования речевой деятельности.

Мы утверждаем, что, по крайней мере, некоторые частицы и вводные слова служат для управления пониманием сообщения адресатом: подтверждают его ожидания и импликатуры или, напротив, сообщают об их нарушениях. При этом поведение участников события моделируется в рамках так называемого «интерактивного подхода», который предполагает учет взаимодействия между говорящим

и слушающим во время речепорождения (Эрвин-Трипп 1975). Существуют различные версии этого подхода, учитывающие те или иные аспекты речи, например, появление ответных реплик, выбор стиля в зависимости от иерархии говорящего и слушающего. Мы обращаем внимание в первую очередь на то, как моделирование взаимодействия предполагает прогнозирование говорящим понимание слушающим сообщения. При выборе средств выражения необходимого ему смысла говорящий сравнивает возможные варианты (синонимы, трансформации) и выбирает тот, который, с его точки зрения, будет понят наиболее легко и безошибочно (Борисова 2006). Это предположение является выводом из постулатов Грайса.

Необходимость выбора из нескольких вариантов (а не, к примеру, употребление единственно нужного слова) возникает потому, что никакой смысл не совпадает с суммой значений языковых средств. Практически всегда от слушающего требуются выводы (импликатуры), позволяющие приблизиться к замыслу говорящего. В тех случаях, когда один из вариантов явно превосходит прочие по близости к замыслу, ситуация выбора отходит на задний план. Но в других ситуациях выбор имеет место и может осознаваться носителем, например, как «муки слова». В любом случае импликатуры разного вида всегда сопровождают любое общение. Иногда такие импликатуры конвенциализуются. Например, вопрос:

(1) *Вы не знаете, который час?*

имеет импликацией вывод: ‘Если знаете, то скажите’.

Фактически этот вывод конвенциализовался , и вопрос воспринимается как косвенный речевой акт просьбы: *Скажите, который час*.

В других случаях выводы могут быть менее очевидными, их можно не сделать, не догадавшись или не зная всех обстоятельств, нужных для выводов. Например, фраза

(2) *Давайте дело делать, а не болтовней заниматься*

может в определенных исторических условиях служить косвенным призывом не участвовать в публичных акциях и т. п. Однако если не знать этих условий, призыв выглядит воспроизведением банальной истины, т. к. возможные выводы, скорее всего, сделаны не будут.

Для того чтобы понимание было более полным, нужные выводы сделаны, а ненужные избегнуты, говорящий вставляет определенные слова, большинство которых и относят к дискурсивным.

Мы предполагаем, что такие корректирующие вставки требуются при привлечении внимания к речевому акту в целом или к отдельному сообщению, в случае перераспределении внимания между отдельными частями сообщения. Это управление вниманием слушающего , что тоже играет определенную роль в управлении пониманием сообщения. Кроме того, использование определенной лексики необходимо и непосредственно при коррекции понимания, которое может быть неверным или неточным в связи с неправильно сделанными выводами или вообще не сделанными.

2. Управление говорящим вниманием слушающего

В первую очередь отметим управление вниманием слушающего. Распределение внимания — важный момент порождения текста, поскольку построение линеаризованного текста предполагает проведение какой-либо важнейшей идеи, относительно которой вся упоминаемая информация играет в той или иной степени вспомогательную роль: служит для иллюстрации, для доказательства основной идеи и т. п. Язык имеет немало средств маркировки информации разной степени важности. Например, если сообщение поместить в деепричастный оборот, чаще всего оно воспринимается как менее значимое, чем в роли сказуемого. Для нашего исследования важно, что эту же функцию выполняют и дополнительные маркеры: междометия, частицы, вводные слова.

Привлечение внимания к речевому акту: его модальной рамке или содержанию, диктуму — актуально в первую очередь в устном общении, т. к. в письменном такие функции осуществляются самим фактом обращения к тексту (адресат раскрывает книгу, открывает файл и т. п.) или, напротив, более эксплицитно и узуализированно, через конструкции, характеризующие начало текста («*Данная статья посвящена…*», «*Мы намерены затронуть…*» и т. п.). В устном общении также имеются ритуальные формы начала беседы (приветствие, извинение), а кроме того, используются междометия привлечения внимания («Эй!»), показатель новизны сообщения *А* начальное *(«А у нас в квартире гость!»*), свидетельство вспоминания *Да*: (*Да, чуть не забыл)*, а также средства, обычно употребляемые при переходе к какой-либо информации, отклоняющейся от основного изложения: *кстати, между прочим, вот*.

Привлечение внимания к отдельным фрагментам сообщений носит характер выделения. А выделение обычно вызвано нарушением каких-либо ожиданий. Например, выделение слова частицей *Да* может быть связано с отрицанием утверждения:

(3) *Да я не сплю!*

Привлечение внимания может вызываться и нарушением ожидаемого порядка изложения: обращения к неожиданной теме, например, вследствие приведения факта в качестве причины:

(4) *Надень шапку — наверху-то холодно,*

в качестве примера:

(5) *Вот твой друг бы этого не сделал*

к уже известному факту (потому что он явился причиной чего-либо) и т. п.

Отметим, что помимо привлечения внимания используются и средства, которые можно было бы назвать средствами отвлечения внимания,

т. е. указания говорящего, что слушающий не должен проявлять интереса и надеяться на получение информации более подробной, чем то, что уже сказано:

(6) *Мы тут посовещались*
(Борисова 1998)

Наконец, видимо, можно говорить и о перераспределении внимания. Например, частица *и* перед сказуемым нередко показывает, что оно не несет, как ожидается, наиболее значимой информации, например:

(7) *Поэтому-то мы и поехали на Юг*

Высказывание уместно в контексте, где факт поездки на Юг не является новостью. (Это понимание роли частицы *и* соответствует ее отмечаемому значению верификации, т. е. установлению истинности предположения. Подробнее значения частицы *и* разобраны в Широкова 1982)

В некоторых случаях управление вниманием является в то же время и управлением пониманием: привлечение внимания к утверждению может быть основанием для возникновения ожидаемых выводов. Например, привлечение внимания при помощи частиц — *то* или *вот* при выделении объекта может быть основанием для противопоставления:

(8) *Учится-то он хорошо, а вот родителей не слушает.*

3. Управление пониманием сообщения: основные виды коррекции (не)понимания

Говоря об управлении пониманием, мы рассматриваем коррекцию возможных импликатур адресата. Имеются и другие способы влияния говорящего на процесс понимания сообщения адресатом, в частности, связанные с корректировкой предполагаемой оценки говорящим события:

(9) *Все пришли в коротеньких юбочках. — Нахалки! — Да вы что, я как раз похвалить их хотела.*

Непонимание было бы предотвращено при своевременной маркировке оценки, например:

(10) *Пришли, умнички, в коротких юбочках.*

Большое место в коммуникации занимают и способы предотвращения и исправления ложной референции. Однако здесь такие способы корректировки понимания рассматриваться не будут.

По большей части корректировки понимания путем уточнения или предотвращения импликатур слушающего осуществляются наречиями и словосочетаниями, называемыми вводными словами, а также некоторыми частицами.

Помимо ссылочно-метатекстовой функции (*по-моему, во-первых*) и функции указания на модальность (*конечно, наверное*) вводные слова способны маркировать более тонкие нюансы управления пониманием. В частности, в кандидатской диссертации Е. С. Яковлевой (Яковлева 1983) отмечена возможность использования ряда вводных слов (в частности, *буквально*) для эмфазы, т. е. подчеркивания высокой степени признака:

(11) *Он буквально замучил меня.*

Здесь наречие ведет себя как частица *просто*, т. е. служит для подчеркивания высокой степени свойства, выраженного глаголом (свойство «мучительности», входящее в значение этого слова). Но и в этом случае подчеркивание построено на коррекции ряда возможных выводов, которые очень приблизительно можно описать следующим образом: говорящий полагает, что очень высокая степень (фрагмент значения слова *замучить*) может восприниматься как преувеличение, поэтому уточняет, что его нужно воспринимать буквально, т. е. точно.

(Поскольку выражение высокой степени требует постоянного усиления, значение наречия — или уже частицы — *буквально* постепенно приблизилось к маркировке гиперболы, что, как отмечает Е. С. Яковлева, является интересным случаем энантиосемии, см. Яковлева 1983).

4. Коррекция импликатур и пресуппозиций слушающего с помощью дискурсивных слов

Слова, связанные с маркировкой неверных ожиданий и выводов слушающего, представляют собой группу, объединяющую не синонимичные, но нередко функционально взаимозаменимые наречия и словосочетания. Это: *по правде говоря, честно говоря, вообще, вообще-то, собственно, фактически, в сущности, по сути, на самом деле* и некоторые другие. Рассматриваемые нами употребления в роли корректоров понимания текста — не единственные значения большинства этих слов и словосочетаний. Почти все они могут вносить и дополнительное значение, в том числе синтаксическое: уступительности

(12) *Вообще-то не положено, но мы пойдем навстречу,*

а также модальные оттенки сожаления и даже извинения:

(13) *Честно говоря, я надеялся, что обойдется*

и ряд других.

Как мы видим, внутренняя форма рассматриваемых слов практически всегда связана с истинностью и обобщением.

О корректирующей функции двух слов из данного списка говорили И. М. Кобозева и Н. С. Латышева (Щедрина). Они рассмотрели слова (которые они называют «дискурсивные единицы») *собственно* и *фактически* и отметили, что эти слова могут употребляться в основном значении, а также для выполнения ряда дискурсивных функций, связанных с корректировкой понимания сообщения.

И. М. Кобозева и Н. С. Латышева (Щедрина) обратили внимание на свойство слова *собственно* — отмечать основную, ядерную часть понятия. Это значение реализуются в таких примерах как:

(14) *Молодое поколение уже выбирает не клипы, не движущиеся картинки, а собственно кино. [«Экран и сцена» 2004г. (НКРЯ)*

Это же значение встречается и в терминосистемах, где подкласс нередко имеет то же название, что и включающий его класс: *собственно круглые черви* (как часть класса круглых червей, включающих разновидности и с другими названиями). Заметим, что в таких случаях перед нами не вводное слово, а наречие, что отражают авторы, не выделяющие слово запятыми. И по многим функциям это слово близко местоимению *сам*, ср.

(15) *Меня интересуют собственно компьютеры, а не комплекты с периферией*, что можно заменить на

(16) *Меня интересуют сами компьютеры…*

Переход к дискурсивному употреблению авторы статьи связывают с противопоставлением чего-то частного, ядерного более общему. Не вдаваясь в предложенные ими истолкования различных случаев (часть которых можно принять, а с другими не согласиться), заметим, что чаще всего можно говорить о корректировке выводов адресата, связанных с обобщением:

(17) *Да после окончания лечения проблем, собственно, и не было. [Наши дети: Подростки (2004)] [НКРЯ]*

Говорящий подозревает, что адресат предполагает наличие проблем как вывод из предыдущего текста и предупреждает его.

Аналогичные действия предполагаются и при употреблении (в дискурсивной функции) слов *в сущности, по сути*:

(18) *Та же, в сущности, ситуация имела место в полузабытую эпоху аналоговых записей на виниловых пластинках и компакт-кассетах. [Андрей Анненков. Против взлома нет приема. Арестует ли ФБР американского ученого (2002) // «Известия», 2002.11.13]. [НКРЯ]*

В данном примере употребление единицы *в сущности* позволяет автору отмести возникающие (возможно) у адресата выводы об отличиях ситуации «в полузабытую эпоху» от нынешней: выводы представлены как не имеющие отношения к самому главному.

Обращение к сути, т. е. к самому главному в ситуации позволяет противопоставить все возможные выводы адресата тому, что имеет в виду именно автор, который и объявляет сутью то, что представляется верным лично ему.

Дискурсивные функции слов *собственно, в сущности, по сути* получают дальнейшее развитие: возникает возможность усилить вопрос или снизить его вопросительную силу. Пример:

(19) *А чего, собственно, вы хотите?*

может иметь и вид очень настоятельного вопроса, и риторического (последнее отмечено И. М. Кобозевой и Н. С. Щедриной). В этих употреблениях слова выполняют ровно те же функции, что и некоторые частицы , ср.

(20) *Чего же вы хотите?.*

Возникают и некоторые модальные оттенки, в частности, извинения: (21) *Я, собственно, ни на что и не рассчитывал*.

Извинение в этом случае, как и при других употреблениях единиц описываемого типа, является следствием предотвращения вывода слушающего, причем такого, который может содержать оттенки возмущения, обиды и т. п. В частности, в этом примере опровергается вероятный вывод слушающего о каких-то планах говорящего, расцениваемых как слишком нахальные, или чем-то подобном.

При этом *собственно, в сущности, по сути* сохраняют близость, но далеко не всегда являются взаимозаменяемыми. Так, нельзя в значении извинения употребить фразу *Я, по сути, ни на что не рассчитывал*. А примеры

(21) *Он, в сущности, неплохой малый* и

(22) *Он, собственно, неплохой малый* явно различаются по смыслу: *собственно* предполагает большую степень противопоставления.

Коррекция возможных предположений адресата осуществляется и путем прямого противопоставления их реальности (точнее, тому, что представляет реальностью говорящий) при употреблении слов *фактически* и *на самом деле*. Это близко к сопоставлению с сущностью, но носит явно более категорический характер:

(23) *В последние годы налоговые органы фактически превратились в карательные.*

Если апелляция к сути допускает разницу мнений, то против факта не поспоришь, что и делает эти слова манипулятивными, как отмечено в статье И. М. Кобозевой и Н. С. Щедриной.

5. Коррекция импликатур и противопоставление общего и частного

Слова *вообще, вообще-то, вообще-то говоря* явно апеллируют к противопоставлению чего-либо общего и какой-то конкретной данности. Как отмечено в словаре русского языка под ред С. И. Ожегова (23 изд, 1990), слово *вообще* означает «в большей части случаев» и может употребляться при выделении и противопоставлении. В этом значении слово употребляется при расширении сферы действия утверждения:

(24) *А я смотрю кино, в котором она играла, или вообще кино того времени и не могу себе представить, что в том мире, в том времени, когда снималось это кино, был такой же воздух, мухи летали, была пыль в воздухе, и воздух был вот такой, неплотный. [Евгений Гришковец (Пример из НКРЯ).*

(25) *Если вы помните, в православной и вообще в христианской религии дело обстоит именно так. [Ordinamenti (2004) // «Экран и сцена», 2004.05.06] (Пример из НКРЯ).*

Значение расширения, обобщения связано и с дискурсивной функцией перехода к дополнительному материалу (что наблюдается и у слов *кстати, между прочим*):

(26) *Ещё до начала собственно исследования, по предложению Д. Б. Эльконина была проведена серьёзная предварительная работа по знакомству с семьями учащихся, с их родителями, с условиями их жизни, а также получены характеристики детей от учителей младших классов (поскольку изучаемый V класс был сформирован из двух IV). Эта работа дала возможность получить некоторое представление о ребятах, которых нам предстояло изучать, но только потом мы поняли, как приблизительны были эти впечатления.* ***Вообще*** *надо сказать, что, контактируя с родителями в процессе работы, мы пришли к неожиданному выводу. [НКРЯ].*

В приведенном примере вводное слово *вообще* позволяет перейти не к обобщению, а к дополнительной информации о получении неожиданного вывода. Здесь использование этого слова мотивируется тем, что авторы отходят от конкретного случая, показывая, что утверждение может распространяться на более широкие сферы, в частности, на обнаруженный вывод, лежащий за пределами рассматриваемого ранее.

В функции коррекции выводов адресата *вообще* показывает, что возможный вывод из события или текста неточный, преждевременный:

(27) *Как вы его терпите? — Он вообще неплохой. Просто бывают странности.*

Функцию коррекции здесь можно описать так: вывод из частного случая неверен, надо рассмотреть более общую ситуацию.

Более четко функция коррекции заметна у дискурсивных единиц *вообще-то* и *вообще говоря*. *Вообще-то* часто употребляется для противопоставления (чему способствует частица *–то*, имеющая и противопоставительное значение). Помимо противопоставления непосредственно в тексте:

(28) *Вообще-то она вела географию, но тут ей пришлось заменять историка*

это слово помогает противопоставлять и возможные выводы адресата тому утверждению, которое говорящий считает истинным:

(29) *Что нужно региональной компании для того, чтобы обойти столичных системных интеграторов и победить в крупном региональном тендере? Вообще-то уральские компании довольно успешно решают этот вопрос. АКСИТ лоббирует их интересы, и руководители органов власти, как правило, готовы поддержать местных разработчиков. В то же время порой местные компании выигрывали тендеры на поставку оборудования и программного обеспечения в другие регионы [НКРЯ].*

Здесь вопрос, заданный автором, предполагает один из вполне вероятных выводов: «уральские компании имеют проблемы с крупными тендерами». Употребив слово *вообще-то* автор показывает , что вывод может быть опровергнут. И в то же время сохраняется возможность для дальнейших противопоставлений (успешно решают, но иногда проигрывают), предупреждающих возможные выводы адресата (например, вывода «если все хорошо, то зачем говорить»).

6. Выводы

Мы рассмотрели далеко не все дискурсивные слова, используемые для коррекции возможных выводов адресата. И в рамках этих рассмотрений только затронули основные дискурсивные функции, развивающиеся у наречий в функции вводных слов. Несомненно, управление пониманием при помощи этих единиц (а также функционально близких им частиц) не сводится к отмеченным нами управлением вниманием и импликатурами адресата. Однако представляется, что и эти функции должны быть выявлены и проработаны при создании модели, отражающей взаимодействие участников общения. В частности, рассмотренные примеры показывают необходимость учета в семантическом описании не только эксплицитно выраженного смысла, но и наиболее вероятных пресуппозиций и импликатур.

Литература

1. *Борисова Е. Г.* Управление вниманием говорящего при помощи частиц // Труды Международного семинара «Диалог-98» по компьютерной лингвистике и ее приложениям, Казань, 1998 г — сс.84–88.
2. *Борисова Е. Г.* Интерактивный подход в языке. Границы применимости// Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции «Диалог-2006». Бекасово, 2006. — с.82–84
3. *Кобозева И. М., Щедрина Н. С.* Дискурсивные единицы собственно и фактически как средство манипуляции представлениями адресата // Эффективность коммуникации. Сб. статей М.: МГПУ, 2009 — с.95–101.
4. *Широкова Е. Г.* Частица И и некоторые функции усилительных частиц // Семантика служебных слов. сб. науч. тр. / Пермск. гос. ун-т. Пермь, 1982. — С. 166–176.
5. *Эрвин-Трипп С. М.* Язык. Тема. Слушатель. Анализ взаимодействия
6. *// Новое в лингвистике. Вып. VII. Социолингвистика. — М., 1975. — С. 336–362*
7. *Яковлева, Е. С.* Значение и употребление модальных слов, относимых к разряду показателей достоверности/недостоверности: автореф. дис. . .канд. филол. наук / Яковлева Е. С. М., 1983.
8. *Abraham W.* (ed.) Discourse Particles: Descriptive and Theoretical Investigations on the Logical, Syntactic and Pragmatic Properties of Discourse Particles in German and English. Pragmatics & Beyond New Ser., 12,1991.
9. *Rathmayr R.* Die russischen Partikeln als Pragmalexeme. München, 1985.

DATA-DRIVEN SPEECH PARAMETER GENERATION FOR RUSSIAN TEXT-TO-SPEECH SYSTEM

Chistikov P. G. (chistikov@speechpro.com),
Korolkov E. A. (korolkov@speechpro.com)

Speech Technology Center ltd, St. Petersburg, Russia

We propose a speech parameter generation approach for Russian based on hidden Markov models. The speech parameter sequence is generated from HMMs whose observation vectors contain speech characteristics. As a baseline we use the spectrum represented by mel-frequency cepstral coefficients (MFCC), pitch and duration parameters. All of them can be easily complemented by any other parameters, improving the quality. For the creation of the voice model we use linguistic and prosodic features which are the observations of every allophone in the utterance. This paper also presents the results of research into selecting the most effective features to characterize an allophone. Experimental results show that Russian speech can be successfully parameterized and an arbitrary utterance can be synthesized from the generated parameters.

Key words: speech processing, speech synthesis, text-to-speech system, tree based clustering, speech parameterization

1. Introduction

Voice analysis and synthesis have been vastly studied in recent years. Many applications and methods have been developed in these areas. The increasing availability of large speech databases makes it possible to construct speech synthesis systems by means of a data-driven or corpus-based approach, by applying statistical learning algorithms. These systems can be automatically trained and are able not only to produce natural and high-quality synthetic speech but also, importantly, to imitate voice characteristics of the original speaker.

For constructing such a system the use of hidden Markov models (HMMs) is widely employed [1–3]. HMMs have been successfully applied to model the sequence of speech spectra in speech recognition systems. The quality of HMM-based speech recognition systems has been improved by techniques which utilize the flexibility of HMMs. These are mixtures of Gaussian densities, tying mechanism, context-dependent modeling dynamic feature parameters and speaker and environment adaptation methods. By applying these techniques to a TTS system we can also support the contemporary tendency towards the synthesis of voices with different styles and emotions [4–5].

In this research we apply the HMM-based approach to speech parameterization to Russian. In our work we use spectrum, pitch and duration as baseline parameters which can be easily complemented by dynamic features and any other features to improve the quality of synthesized speech. It is worth noting that the voice model can be created with no need for large databases [6–8].

The paper is organized as follows: in Section 2 all the procedures that are carried out by the TTS system are described, from the source database to the synthesized speech. Section 3 describes the voice model building methods to synthesize a given utterance. In Section 4, linguistic and prosodic features are presented. Section 5 deals with the system implementation and experimental results. The conclusions are given in Section 6.

2. TTS Engine description

The speech generation mechanism is a subpart of the fully-functional text-to-speech system which has been constructed on the basis of the Vital Voice [9] system and HTK [10]. The procedures of voice model building and synthesis of an arbitrary utterance by the speech synthesis engine are shown in Figure 1.

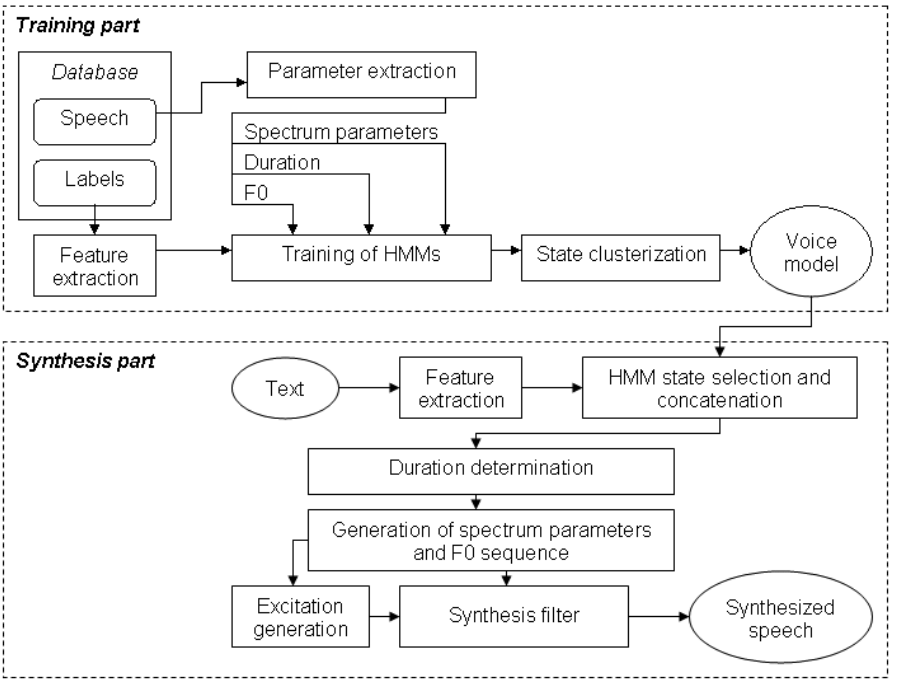


Fig. 1. Diagram illustrating the basic steps conducted by the speech synthesis engine

2.1. Training

The training starts with parameter extraction. For each allophone from the database, linguistic and prosodic features and voice parameters are calculated. Then context-dependent HMMs are trained. In the last step, HMM states are clustered based on linguistic and prosodic features. Eventually we have the voice model. Section 3 gives these steps in detail.

2.2. Synthesis

The synthesis procedure starts with transcribing the utterance to an allophone sequence and generating the linguistic and prosodic features which are eventually used to select corresponding leaves from each of the 2N+1 (where N is the number of states in HMM) decision trees generated by the context-clustering procedure in the training step. At the end of this stage, three logical HMM sequences, whose states correspond to the selected leaves, are obtained.

The three above-mentioned HMM sequences are used to derive spectrum parameters, fundamental frequencies and state durations. Then excitation is generated based on F0 and energy, and synthesized speech is extracted by the synthesis filter based on excitation and spectra [11–13].

3. Voice model building

Initially a sequence of fundamental frequencies $\{F0^1, \ldots, F0^K\}$, including voicing decision information (if F0 is 0 then the frame is considered unvoiced), where K is the total number of frames of all sentences from the training database, is extracted on a short-term basis. Simultaneously, a sequence of spectrum parameter (mel-cepstral coefficients [14]) vectors which represent speech envelope spectra, $\{c^1, \ldots, c^K\}$, is obtained. Each MFCC vector $\overline{c}^i=[c_0^i \ldots c_M^i]$, where the i indicates the frame number and $[\cdot]^T$ means transposition, is derived through an M-th order mel-cepstral analysis.

After that, linguistic and prosodic features for each allophone of all the sentences of the training database are estimated. The description of linguistic and prosodic features for Russian is given in Section 4.

In the next step HMM prototypes for each allophone are created. Each HMM corresponds to a no-skip N-state left-to-right model with N = 5. Each output observation vector $\overline{o}^i$ for the i-th frame consists of 2 streams, $\overline{o}^i=[\overline{o}_1^{iT}, \overline{o}_2^{iT}]^T$, where stream 1 is a vector composed by MFCCs, their delta and delta-delta components; and stream 2 is a vector composed by F0s, their delta and delta-delta components as well.

The observation vector $\overline{o}^i$ is the output of an HMM state n according to a probability given by

$$\beta_n(\overline{o}^i) = \prod_{j=1}^{2}[\sum_{l=1}^{R_j} \omega_{njl} \mathrm{N}(\overline{o}_j^i; \mu_{njl}, \Sigma_{njl})], \quad (1)$$

where $\overline{o}_j^i$ means a Gaussian distribution with mean vector μ and covariance matrix Σ, ω_{njl} is the weight for the l-th mixture component of the j-th stream of vector $\overline{o}_j^i$ (output of the state n) with R_j being the corresponding number of mixture components. The stream vectors are modelled by single-mixture continuous Gaussian distributions where the dimensionality is 3(M+1) for $\overline{o}_1^i$ and 3 for $\overline{o}_2^i$.

For each k-th HMM the durations of the N states are considered as vectors $\overline{d}^k=[\overline{d}_1^k,...,\overline{d}_N^k]^T$, where $\overline{d}_n^k$ represents the duration of the n-th state. Furthermore, each duration vector is modelled by an N-dimensional single-mixture Gaussian distribution. The output probabilities of the state duration vectors are thus re-estimated by Baum-Welch iterations in the same way as the output probabilities of the speech parameters [15].

During the voice model building, a tree-based clustering technique is applied to the HMM-states of MFCC and F0 values, as well as to the state duration models. In the end of the process, 2N+1 different acoustic decision trees are generated: N trees for mel-cepstral coefficients, N trees for F0 features and finally one tree for state duration.

4. Contextual features

When speech synthesis systems are constructed, some parameters are necessary to provide a natural rendering of the prosody. These parameters might include context-dependent items, for instance preceding/following phone, syllable, word, sentence etc [17].

The determination of contextual features for a particular language is based on linguistic and prosodic parameters. Apart from this theoretical approach, empirical analysis can also be implemented in order to tune the features by extending the factors that are important and eliminating the ones which are not [18].

The contextual features listed in Table 1, which were selected as the most informative ones to build the voice model, were first obtained from those used in HMM-based Brazilian Portuguese speech synthesis [18] and eventually adjusted through theoretical and empirical methods to Russian.

5. Implementation

5.1. The corpus

In this work we used the speech corpora (one male and one female voices) initially prepared for the Vital Voice TTS system [9]. They correspond to six hours of speech excluding silence regions. The databases were recorded at a sample rate of 22 050 Hz with 16 bits per sample.

The phonetic labeling of the database was carried out using the phone set described in [16]. Time label boundaries were obtained manually. Further, syllable and word labeling were also manually conducted for each sentence.

Table 1. Contextual features

Allophone features	
Phone before previous	Phone after next
Previous phone	Phone position from the beginning of the syllable
Current phone	Phone position from the end of the syllable
Next phone	
Syllable features	
Previous syllable	Syllable position from the end of the word
Current syllable	Syllable position from the beginning of the sentence
Next syllable	Syllable position from the end of the sentence
Number of phones in the previous syllable	Number of stressed syllables before current syllable in the sentence
Number of phones in the current syllable	Number of stressed syllables after current syllable in the sentence
Number of phones in the next syllable	Vowel name in the current syllable
Syllable position from the beginning of the word	
Word features	
Part of speech of the previous word	Number of syllables in the current word
Part of speech of the current word	Number of syllables in the next word
Part of speech of the next word	Word position from the beginning of the sentence
Number of syllables in the previous word	Word position from the end of the sentence
Sentence features	
Number of syllables in the current sentence	Sentence type (declarative, exclamation etc.)
Number of words in the current sentence	

5.2. Parameter extraction

Fundamental frequencies and mel-cepstral coefficients were extracted from the speech corpus in every 10ms frame. MFCCs were obtained through a 25-th order analysis (M=25) by means of a 25ms Henning window.

5.3. Generated decision trees

Figures 2 and 3 show the top part of the decision tree for mel-cepstral coefficients of [a1] phoneme, and state durations, respectively. By observing Figure 2 and Figure 3, and assuming that top nodes are more important when selecting the parameter which is being clustered, we can notice that the information about the syllable, word and sentence is more crucial for state duration. On the other hand, questions related to phonemes are more significant to the tree of mel-cepstral coefficients.

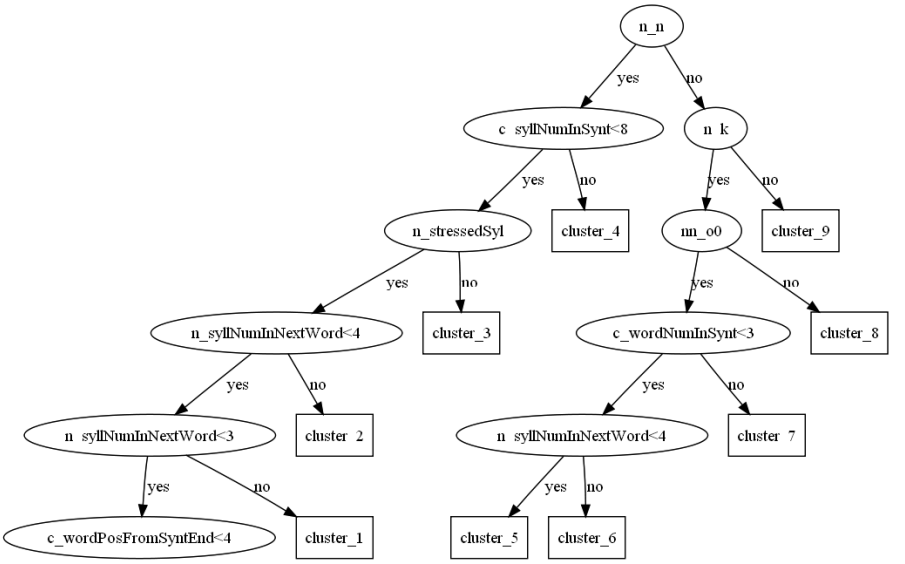


Fig. 2. Top of the decision-tree constructed to cluster the third HMM state for mel-cepstral coefficients of [a1] phoneme

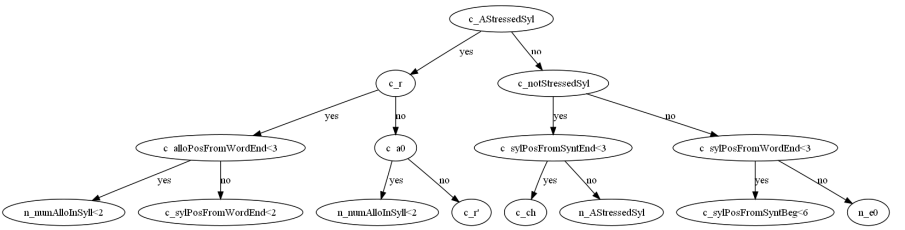


Fig. 3. Top of the decision-tree constructed to cluster the distribution of state durations

5.4. Example of speech synthesis

Figure 4 presents the spectrograms for the natural sentence “Это моя сестра” (this is my sister) and its synthesized version. It should be noted that the utterance is not part of the training database.

Aside from the reproduction of phones it can be also observed from Figure 4 that the synthesized version presents a speaking rate similar to that of the natural speech. This shows an important characteristic of the HMM-based speech synthesis approach: the ability to imitate the prosody of the speech corpus which was used to build the voice model.

Although it has been reported that even with a small database it is possible to synthesize speech, lack of data strongly affects the quality of synthesized speech. Once the HMMs do not properly reflect the characteristics of some linguistic and prosodic feature sets, inconsistent parameters may be generated during synthesizing, and as a result we could have poor quality speech. Thus the process of preparing the database is very important for building appropriate voice models.

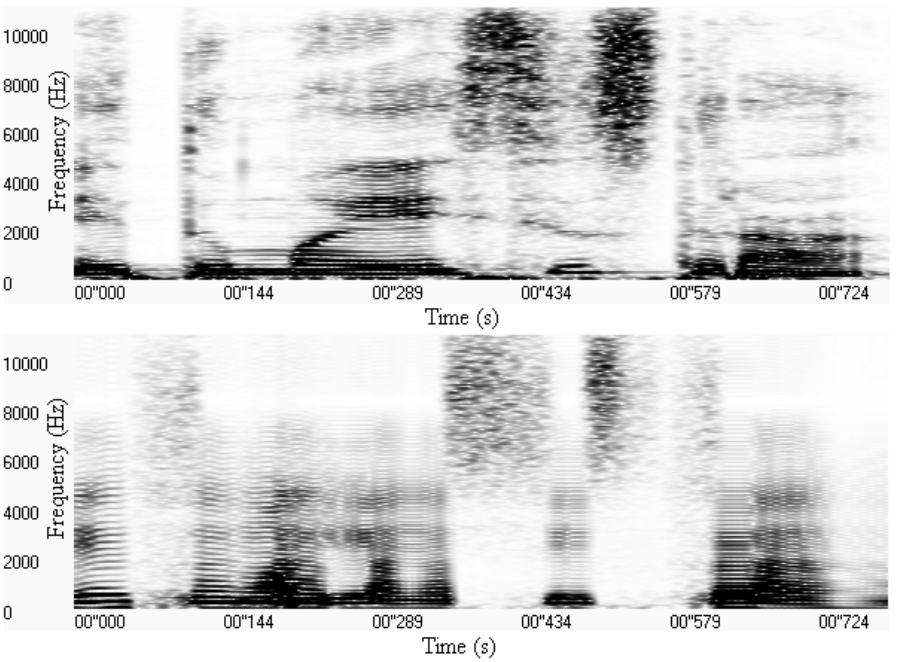


Fig. 4. Spectrograms for the natural sentence “Это моя сестра” (this is my sister) (top) and its synthesized version (bottom)

6. Conclusions

This paper described an approach for corpus-based speech parameter generation for a Russian text-to-speech system. The engine is based on a method where the speech parameters are obtained from HMMs whose observation vectors consist of spectum, F0 and duration features. We performed context-clustering to achieve a greater flexibility of the algorithm and to make it possible to use the voice model even when a small database is used. The clusterization procedure is based on linguistic and prosodic features of Russian which were also presented in this paper. Experimental results show that Russian speech can be successfully parameterized and any utterance can be synthesized from the generated parameters.

References

1. *Black A.* Unit selection and emotional speech. Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH). 2003.
2. *Black A., Taylor P., Caley R.* The Festival Speech Synthesis System, http://www.festvox.org/festival.
3. *Campbell N.* Towards synthesizing expressive speech; designing and collective expressive speech data. Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH). 2003.
4. *Fukada T., Tokuda K., Kobayashi T., Imai S.* An adaptive algorithm for mel-cepstral analysis of speech. Proceedings ICASSP-92. 1992, pp. 137–140.
5. *Fukada T., Tokuda K., Kobayashi T., Imai S.* An adaptive algorithm for mel-cepstral analysis of speech. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 1992.
6. *Imai S., Sumita K., Furuichi C.* (1983) Mel log spectrum approximation (MLSA) filter for speech synthesis. IECE Trans. A, Vol. J66-A, no.2, pp.122–129.
7. *Korolkov E, Glavatskih I, Talanov A.* Vital Russian voice text-to-speech system based on unit selection method. Proceedings of the 36th International Philological Conference on formal methods for the analysis of the Russian language. Russia, 2008.
8. *Maia R., Zen H., Tokuda K.* An HMM-based Brazilian Portuguese Speech Synthesis and Its Characteristics. Revista da Sociedade Brasileira de Telecomunicações. 2006.
9. *Maia R., Zen H., Tokuda K., Kitamura T., Resende F.* Towards the development of a Brazilian Portuguese text-to-speech system based on HMM. Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH). Geneva, 2003, pp. 2465–2468.
10. *Shichiri K., Sawabe A., Yoshimura T., Tokuda K., Masuko T., Kobayashi T., Kitamura T.* Eigenvoices for HMM-based speech synthesis. Proceeding of the International Conference on Spoken Language Processing (ICSLP). 2002.
11. *Smirnova N., Chistikov P.* Software for Automated Statistical Analysis of Phonetic Units Frequency in Russian Texts and its Application for Speech Technology Tasks. Proceedings of the Dialogue-2011 International Conference. 2011.

12. *Tokuda K., Kobayashi T., Fukada T., Saito H., Imai S.* (1991) Spectral estimation of speech based on mel-cepstral representation. IEICE Trans. A, Vol. J74-A, no.8, pp.1240–1248.
13. *Tokuda K., Zen H., Black A.* An HMM-based speech synthesis applied to English. Proceedings of IEEE Workshop in Speech Synthesis. 2002.
14. *Yamagishi J., Tachibana M., Masuko T., Kobayashi T.* Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis. Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 2004.
15. *Yoshimura T., Tokuda K., Masuko T., Kobayashi T., Kitamura T.* Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH). 1999.
16. *Yoshimura T., Tokuda K., Masuko T., Kobayashi T., Kitamura T.* Speaker interpolation in HMM-based speech synthesis. Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH). 1997.
17. *Young S., Evermann G., Gales M.* (2006) The HTK Book (for HTK Version 3.4).
18. *Zen H., Tokuda K., Masuko T., Kobayashi T., Kitamura T.* Hidden semi-Markov model based speech synthesis. Proceedings of the International Conference on Spoken Language Processing (ICSLP). 2004.

НКРЯ КАК ИНСТРУМЕНТ СОЦИОЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ. ЭПИЗОД IV. ПОЛ ГОВОРЯЩЕГО И ДЛИНА РЕПЛИКИ

Даниэль М. А. (misha.daniel@gmail.com)
НИУ ВШЭ, МГУ, Москва, Россия

Зеленков Ю. Г. (yuryz@yandex.ru)
Яндекс, Москва, Россия

В статье обсуждается использование Национального корпуса русского языка для социолингвистических исследований. Предлагается использовать имеющуюся в Устном корпусе разметку по полу говорящего в качестве основы для статистического анализа различий между речью мужчин и женщин. Подробно обсуждаются результаты одного направления анализа — сравнительная длина высказываний (реплик) мужчин и женщин, измеренная в количестве словоформ. Обнаружена статистическая значимость различий по этому параметру между мужчинами и женщинами в публичной коммуникации — мужчины говорят больше женщин — при отсутствии статистически значимых различий в частной коммуникации. Также значимым оказывается пол адресата, причем корреляция снова носит существенно разный характер в подкорпусе публичной и частной речи. В частной коммуникации реплики говорящего оказываются несколько длиннее в разговоре с человеком того же пола; в публичной коммуникации реплики говорящего оказываются существенно длиннее при обращении к женщине, чем к мужчине, вне зависимости от пола говорящего. Эти выводы хорошо согласуются с результатами гендерных исследований в западной социолингвистике: различия между мужской и женской речью не абсолютны, но зависят от ситуации общения; публичный дискурс является ареной языкового соревнования, в котором преимущество остается за мужчинами, в то время как в частной коммуникации соревновательность исчезает.

Ключевые слова: гендер, корпусная лингвистика, социолингвистика, русский язык, Национальный корпус русского языка, речевое взаимодействие

Как показывают исследования, мужчины считают, что женщины много говорят, не потому, что женщины говорят больше мужчин, а потому, что они вообще открывают рот
Социолингвист, пожелавший остаться анонимным

Введение

В недавней статье, представленной в рамках конференции «Диалог», один из ведущих российских социолингвистов В. И. Беликов обосновывает точку зрения, согласно которой НКРЯ не может служить надежным источником данных для социолингвистического анализа (Беликов 2011). Основным аргументом для В.И. является нерепрезентативность НКРЯ. Мы могли бы добавить многие другие (как нам кажется, более бесспорные) недостатки (устного) Корпуса как источника социолингвистических данных: например, его ориентированность на лексико-грамматические характеристики, что делает невозможным проведение исследований в одном из самых интересных и популярных на западе направлений социолингвистики — в сфере социофонетики. Обращает на себя внимание, однако, следующее уточнение: на социолингвиста НКРЯ «не рассчитан, поскольку ни в коей мере не может считаться сбалансированным с точки зрения решаемых им задач». Очевидно, что это утверждение может быть верно только в зависимости от того, как мы определим круг задач и интересов социолингвиста. В том разделе цитируемой статьи, которая посвящена социолингвистике, В. И. Беликов убедительно критикует некоторые неудачные примеры анализа социальной вариативности в сфере лексики. В настоящей статье, однако, мы хотим на примере другой социолингвистической задачи показать, что по крайней мере Устный корпус, несмотря на существенные недочеты, является социолингвистическим ресурсом с большим исследовательским потенциалом.

Следует сразу оговориться, что речь пойдет об оффлайновом исследовании текстов НКРЯ. Корпус выступает здесь не в качестве интерактивного инструмента (а именно примеры таких исследований рассматривает В. И. Беликов), но как коллекция аннотированных текстов; а в таком виде он недоступен простому пользователю. Тем не менее, главный тезис остается в силе: с нашей точки зрения, существуют такие социолингвистические задачи, для которых коллекция текстов НКРЯ, вопреки точки зрения, высказанной в (Беликов 2011), является достаточно репрезентативной и позволяет получать интересные результаты.

По состоянию на момент написания статьи Устный корпус включает около 3,000 документов общим объемом почти 10 млн. словоупотреблений. Все документы Устного корпуса разделены на три группы — подкорпус

публичной речи (около 1,000 документов, около 5,1 млн. словоупотреблений), подкорпус устной непубличной речи (около 1,300 документов, более 1,1 млн. словоупотреблений) и подкорпус речи кино (около 690 документов, около 3,8 млн. словоупотреблений). Подробно мотивацию включения речи кино в Устный корпус (проблема спонтанности) см. в (Гришина, Савчук 2009). Важным свойством Устного корпуса является наличие в разметке простейшей информации о говорящих. Для пользователя портала НКРЯ это значит, что поиск можно ограничивать полом и возрастом говорящего. Другая информация об Устном корпусе для задач, обсуждаемых в настоящей статье, может считаться несущественной.

Включение в разметку базовых сведений о говорящем наталкивает на мысль о возможности офлайнового статистического исследования Устного корпуса НКРЯ с социолингвистической точки зрения: например, на предмет различий в речи мужчин и женщин. В настоящей статье по разным причинам представлен лишь один частный результат такого рода исследований — исследование корреляции длины высказывания с полом говорящего, результаты которого в целом соответствуют теоретическим ожиданиям. В заключительной части статьи мы кратко охарактеризуем другие возможные направления социолингвистического анализа Устного Корпуса НКРЯ.

Средняя длина высказывания как социолингвистическая переменная

Хорошо известно, что предметом социолингвистики являются не только собственно социально значимые языковые феномены, но и рефлексия носителей о взаимосвязи социальных и лингвистических категорий. Одним из самых известных стереотипов о женской речи является представление о том, что женщины много говорят. При помощи Устного корпуса этот стереотип легко верифицировать. Нами было посчитано распределение длин мужских и женских реплик. Репликой считалась последовательность высказываний одного говорящего до момента смены говорящего или конца документа. Длиной реплики считалось количество словоупотреблений в реплике. В следующих таблицах показано распределение длин мужских и женских реплик. Общий объем корпуса в репликах составляет 103 037 и 125 642 для мужчин и женщин, соответственно. На Графике 1 показана зависимость между длиной реплики (в словоупотреблениях, ось X) и долей реплик данной длины в корпусе (процентах от общего числа реплик, ось Y). Длинный штрих отражает эту зависимость для мужских реплик, короткий штрих — для женских, а тонкая непрерывная линия — по устному корпусу в целом.

Для всех трех кривых максимальную долю составляют однословные реплики, двух-, трех- и четырехсловные реплики имеют примерно равную долю, а начиная с пяти словоупотреблений доля реплик соответствующей длины

начинает стремительно падать. Однако видно, что кривые для мужских и женских реплик ведут себя различно, так как для длин реплик от одного до десяти словоупотреблений кривая для женских реплик проходит над той же кривой для мужских реплик. Между 11 и 13 словоупотреблениями кривые практически совпадают, а начиная с 14 кривая для женских реплик навсегда уходит «под» кривую для мужских реплик. Для длинных реплик кривые на Графике 1 располагаются визуально близко, однако вес отличия (который можно посчитать, например, как в Графике 2 — отношение разности долей реплик данной длины у женщин и мужчин к доле реплик этой длины во всем Устном корпусе) начиная с 14 словоупотреблений растет.

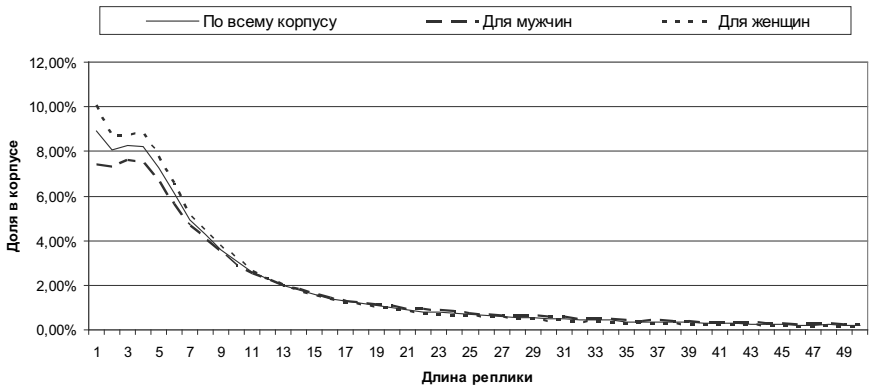


График 1. Зависимость длины реплики и доли в корпусе

График 2. «Взвешенная» разница долей реплик в мужском и женском корпусе (зависимость от длины реплики)

Далее речь будет идти только о средних длинах (без учета распределений). По всему устному корпусу средняя длина реплики мужчины составляет

26,34 словоформы, а средняя длина реплики женщины — 15,41 словоформы. Результат вполне однозначно указывает на то, что обсуждаемый языковой стереотип неверен. Это соответствует результатам, полученным для других языков и на основании иных методов анализа (например, сравнительная длительность речи мужчин и женщин в рамках публичных семинаров, подробнее см. Вахтин, Головко 2004).

Более интересным и содержательным, но также вполне ожидаемым исходя из накопленных в зарубежной социолингвистике результатов, является следующее уточнение. При разделении исходных данных на публичную и частную речь оказывается, что расхождения в длине реплик у мужчин и женщин выглядят для этих двух подкорпусов существенно разным образом.

Таблица 1. Сравнение средней длины мужских и женских реплик по типам текстов (в словоформах)

	мужчины	**женщины**
Устный корпус в целом	26,34	15,41
Публичная речь	30,36	20,65
Частная речь	[10,77]	[10,29]

Из таблицы видно, что различие между публичной речью мужчин и женщин очень значительно (оно приближается к различию по Устному корпусу в целом, так как именно публичная речь составляет его наиболее значительную долю). При этом в частной речи различие в средней длине реплик оказывается статистически незначимым (здесь и далее статистическая значимость определялась по t-критерию Стьюдента на уровне значимости 0,001). Это может быть интерпретировано как подтверждение делаемых в западной социолингвистике обобщений, согласно которым женщины в большей степени ориентированы на частную, чем на публичную речь (Wodak, Benke 1998), а также прекрасно подтверждает один из основных тезисов этого же обзора и общий тренд развития исследований женской речи — различия между речью мужчин и женщин носят не абсолютный, а контекстуальный характер, так что изучение параметров варьирования вообще, в отрыве от коммуникативной ситуации, либо дает вообще неправильные результаты, либо значительно их обедняет.

Иными словам, если в ситуации публичной коммуникации женщины говорят меньше, чем мужчины (или их собеседники не дают им говорить столько же, сколько мужчинам — ср. анализ перебиваний в Tannen 1990, также обсуждаемый в Вахтин, Головко 2004), то в ситуации частного общения они, как кажется, коммуницируют наравне с мужчинами. В этой связи интерес представляет еще один тип калькуляций — относительная длина мужских и женских реплик в зависимости от пола адресата. Ср. следующую таблицу:

Таблица 2. Сравнение средней длины мужских и женских реплик по типам текстов и в зависимости от пола собеседника

Публичная речь		
	пол адресата	
	муж	жен
муж	35.75	46.58
жен	27.28	39.30

Частная речь		
	пол адресата	
	муж	жен
муж	12.94	10.84
жен	9.49	12.23

Из таблицы видна разнонаправленная корреляция между длиной реплики и соотношением полов говорящего и адресата. В частной речи реплики женщин оказываются статистически значимо (хотя и незначительно) длиннее, если они обращены к женщинам. Разница в длине реплик мужчин в зависимости от пола собеседника оказывается еще менее заметной (хотя, опять же, статистически значимой), но демонстрирует ту же корреляцию. Отметим также, что разница между средней длиной реплики женщины, обращенной к женщине, и средней длиной реплики мужчины, обращенной к мужчине, оказывается статистически незначимой, как и разница между средней длиной реплики мужчины, обращенной к женщине, и средней длиной реплики женщины, обращенной к мужчине. Иными словами, в ситуации частной коммуникации говорящий строит более длинные реплики при общении с человеком того же пола; другие различия нивелированы.

Статистическая картина в публичном подкорпусе принципиально иная. Во-первых, зависимость длины реплик от пола собеседника носит гораздо более ярко выраженный характер. Во-вторых, она устроена совершенно иным образом — если средняя длина реплики женщины, обращенной к женщине, превосходит (и значительно) среднюю длину реплики женщины, обращенной к мужчине, то средняя длина реплики мужчины, обращенной к мужчине, не больше, а меньше, причем значительно меньше — средней длины реплики мужчины, обращенной к женщине.

Таким образом, если для зависимости средней длины реплики от пола адресата в частной речи операциональной является категория гендерной гомогенности / гетерогенности (пол адресата отличается от / совпадает с полом говорящего), то в публичной речи определяющую роль играет категория пола адресата (мужской / женский). Или, выражаясь несколько менее формально: в частном разговоре женщина больше говорит с женщиной, а мужчина — с мужчиной, в то время как в публичной речи женский пол адресата значительно увеличивает продолжительность реплики не только у мужчины, но и у женщины.

Известный тезис о соревновательном характере публичного речевого поведения, в котором мужчины преуспевают заметно более женщин (Tannen 1990), согласно полученным нами данным, оказывается применим только к публичному речевому поведению (см. обсуждение гендерных теорий типа power/dominance theories у Wodak, Benke 1998), причем женщины пользуются «гендерной слабостью» адресата так же активно (хотя и менее успешно), как мужчины.

Следует сделать следующую важную оговорку. Как ясно из обсуждения данных, в нашем исследовании контролируемыми переменными являлись только пол говорящего (собственно экстралингвистическая переменная) и общий тип текста согласно классификации НКРЯ (частный или публичный дискурс; о социолингвистическом статусе таких параметров, задающих функционально-коммуникативный контекст, см. все в том же обзоре Wodak, Benke 1998).

Контроль других классических экстралингвистических переменных (возраст, образование, социальный класс) технически невозможен — эти данные просто отсутствуют в разметке НКРЯ. Впрочем, даже если их можно было бы включить, по нашему мнению, они не оказали бы существенного влияния на полученные результаты. Устный Корпус НКРЯ не является репрезентативным ни (наверняка) с точки зрения возраста, ни (с большой степенью вероятности) социального класса говорящих. Судя по жанровому составу Корпуса, можно было бы высказать предположение, что тексты «центрированы» на испытуемых среднего достатка в возрасте от 20 до 40 лет и слабо отражают более маргинальные (с точки зрения типовых социолингвистических исследований) группы (дети, пожилые люди, люмпены и пр.).

С другой стороны, более внимательный контроль типа текста не только возможен, но и абсолютно необходим для продолжения этого исследования, которое можно таким образом признать лишь пилотным. Во-первых, некоторые жанры являются принципиально неженскими (например, парламентские слушания). Впрочем, доля таких текстов в корпусе пренебрежимо мала (менее двух процентов подкорпуса публичной речи).

Во-вторых, что гораздо более важно, многие устные жанры ведут себя с точки зрения длины реплики совершенно особым образом. Так, в Устный Корпус НКРЯ включены встречи с читателями и слушателями, доклады, интервью, спортивные комментарии, лекции, пресс-конференции, рассказы, репортажи (публичная речь), а также рассказы и пересказы (частная коммуникация). Все эти жанры устроены с точки зрения длины реплик иначе, чем, например, токшоу или дискуссии: на одного из говорящих (будем условно называть его «нарратором») приходится большая часть дискурса, в то время как собеседники, если они вообще присутствуют, лишь задают вопросы и вставляют небольшие комментарии. Мы предполагаем, что внутри публичной и частной коммуникации такие «всплески» длительности нивелируются тем, что нарратор в большей части подобных жанров может быть как мужчиной, так и женщиной. Однако сравнение длины реплики в публичном корпусе и в частной коммуникации, на котором мы сделали акцент выше, может от этого сильно страдать (особенно если учесть, что такие особые типы публичных текстов составляют около четверти всего подкорпуса публичной устной речи).

В целом очевидно, что в дальнейшем следует более дробно классифицировать коммуникативные роли говорящих, выделяя по крайней мере роль нарратора. Проще всего просто исключить такого рода данные из анализа. Корреляция между длиной реплики нарратора и его полом составляет предмет для дополнительного исследования, но приходится признать, что «сливать»

этот показатель с длиной реплики в обычной публичной дискуссии, как это сделано в представляемой пилотной версии исследования, в любом случае неправильно — так же неправильно, как сливать длину реплики в частной и публичной коммуникации.

Заключение

Выше было показано, что, несмотря на поверхностность социолингвистической разметки текстов Устного корпуса, оффлайновый статистический анализ текстов позволяет сформулировать нетривиальные интерпретации и обобщения, согласующиеся с результатами, которые были получены иными методами и для других языковых сообществ — например, показать контекстность различий речевого поведения мужчин и женщин, то есть, в частности, перейти от sex-based парадигмы к парадигме, опирающейся на гендер как интерактивную социальную категорию. Конечно, это лишь очень частный результат анализа данных, которые, с нашей точки зрения, являются богатейшим полигоном социолингвистических исследований (причем даже этот частный результат требует дополнительной верификации с учетом более детальной жанровой классификации текстов). Попробуем наметить возможные дальнейшие социолингвистические применения устных текстов НКРЯ.

С точки зрения исследования гендерных различий большой интерес представляют статистически значимые различия между частотностью (частотным рангом) конкретных языковых единиц — лексем, грамматических категорий или их сочетаний, а также, возможно, конкретных словоформ. Эти данные относительно легко получить путем статистического анализа Устного корпуса методами, аналогичными методам автоматической рубрикации документов — причем, что важно, без предварительных гипотез о том, что именно мы ищем (ср. Lakoff 1975), так как ранговые списки таких единиц легко составляются автоматически.

Здесь мы позволим себе сделать краткое отступление от основной темы. Параллельно с данным исследованием одним из авторов статьи разрабатывался метод автоматического определения пола автора высказывания по тексту этого высказывания (в графическом представлении). Эта задача в целом аналогична задаче построения автоматического рубрикатора текстов: были выделены «гендерные маркеры», т. е. языковые признаки принадлежности текста к одному из двух классов (мужские vs. женские тексты). Во многих отношениях эта задача отличается от академической задачи поиска расхождений между женским и мужским «вариантами» языка. Например, социолингвисту могут быть интересны статистически значимые расхождения относительной частотности употребления конкретных языковых единиц (скажем, конкретных лексем), а для автоматического определения пола говорящего, который ориентирован в первую очередь на небольшие тексты и отдельные высказывания, сам факт вхождения той или иной лексемы вряд ли окажется достаточным. Мы упоминаем это исследование лишь потому, что алгоритм

опирается на те же данные, что представленный выше анализ длин реплик, и работает с достаточной степенью надежности и уже внедрен в некоторых приложениях поисковой системы Яндекс. Для определения пола говорящих в Устном корпусе НКРЯ он работает в общем с вероятностью более 0,9 (исключением оказывается речь женщин в публичном подкорпусе — факт, который также вполне можно интерпретировать с точки зрения того, что нам известно о различиях в языковом поведении мужчин и женщин; ср. обзор в Wodak, Benke 1998, по-русски также Вахтин, Головко 2004). Эти результаты указывают на статистическую надежность гендерных интерпретаций данных Устного корпуса и подтверждают целесообразность исследования на том же материале ранговых различий между речью мужчин и женщин. К сожалению, менее лексически ориентированные результаты — например, проверка на материале русского языка тезисов Лакофф об относительной частотности тех или иных синтаксических конструкций — либо вообще не доступны для прямого статистического анализа, либо требуют более тонких косвенных методов оценки.

Важным направлением социолингвистических исследований в области гендерных различий является верификация и интерпретация известного еще с первых работ Лабова тезиса о том, что женская речь по сравнению с мужской в большей степени ориентирована на норму (ср. подробный обзор по этой важной темы в Wodak, Benke 1998). В нашем случае ключевым является вопрос о том, что принять за статистику нормы — с чем сравнивать статистические характеристики устной речи мужчин и женщин. Можно условно считать нормативной статистику корпуса современных газет (сейчас — более 100 млн. словоупотреблений). Несмотря на вполне очевидную условность такого решения, полученные нами предварительные результаты отчасти согласуются с предположениями гендерной теории. Так, количество выделяемых алгоритмом языковых маркеров, отличающие публичную речь женщин от «нормативного» газетного корпуса статистически значимо меньше, чем количество маркеров, отличающих публичную речь мужчин. С другой стороны, аналогичный анализ для частной коммуникации дает обратные результаты, что допускает множество интерпретаций, но в первую очередь требует дополнительной проверки валидности статистического анализа.

Подробное обсуждение этих и других вопросов расширения применения НКРЯ для социолингвистических исследований здесь невозможно в том числе по соображениям объема, поэтому мы оставляем его для отдельной публикации. Единственный момент, который хочется кратко затронуть — это кросслингвистическая применимость предлагаемых методов анализа и сопоставимость их результатов по разным языкам. Единственным условием, необходимым для проведения такого рода исследования на материале других языков, является доступность устных корпусов достаточного объема, обладающих по крайней мере той же разметкой параметров говорящего, что и Устный корпус НКРЯ — указанием пола. Конечно, устные корпуса большого объема — относительная редкость, однако для некоторых языков они уже

существуют. В число таких языков входят не только крупные европейские языки, но и некоторые языки других ареалов — например, Восточноармянский национальный корпус (www.eanc.net). Параллельное применение одного и того же метода статистического анализа к материалу устных корпусов нескольких языков разных структур и ареалов (как лингвистических, так и социокультурных) позволит попытаться ответить на один из самых острых вопросов современной вариационной социолингвистики — в какой степени результаты, полученные ею до сих пор и в подавляющем большинстве случаев опирающиеся на языковые ситуации Западной Европы и других развитых стран (например, ориентация женской речи на норму), носят действительно универсальный характер.

Литература

1. *Lakoff, Robin Tolmach* (1975/2004) Language and woman's place. Oxford: Oxford University Press.
2. *Tannen, Deborah* (1990) You just don't understand. Women and men in conversation. NY: Ballantine Books.
3. *Wodak, Ruth and Gertraud Benke* (1998) Gender as a sociolinguistic variable: new perspectives on variation studies. In Florian Coulmas (ed.) The handbook of sociolinguistics. Oxford: Blackwell. Pp. 127–150.
4. *Беликов, В. И.* (2011) Чего не хватает в «оцифрованном мире» лексикографу и социолингвисту. // Труды международной конференции Диалог 2011. Бекасово.
5. *Вахтин, Н. Б. и Е. В. Головко* (2004) Социолингвистика и социология языка. СПб.: Гуманитарная академия.
6. *Гришина, Е. А. и С. О. Савчук* (2009) Корпус устных текстов в НКРЯ: состав и структура // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История. Стр. 129–149.

ВЫБОР ОПТИМАЛЬНОГО НАБОРА ИНФОРМАТИВНЫХ ПРИЗНАКОВ ДЛЯ КЛАССИФИКАЦИИ ЭМОЦИОНАЛЬНОГО СОСТОЯНИЯ ДИКТОРА ПО ГОЛОСУ

Давыдов А. Г. (davydov-a@speetech.by),
Киселёв В. В. (kiselev-v@speetech.by),
Кочетков Д. С. (kochetkov-d@speetech.by),
Ткаченя А. В. (tkachenia-a@speetech.by)

ООО «Речевые технологии», Минск, Беларусь

Рассматривается вопрос минимизации количества информативных признаков при решении задачи классификации эмоционального состояния диктора по голосу. Доклад состоит из четырех частей. Первая часть содержит обзор информативных признаков характеризующих речь. Показывается, что акустические характеристики голоса могут быть условно разделены на пять категорий: просодические, динамические, фонационные, спектральные и энергетические. Во второй части кратко рассматриваются способы определения минимального эффективного набора информативных признаков. В экспериментальной части рассматривается процедура выбора оптимального набора информативных признаков на основе алгоритма sequential feature selection (SFS) с использованием мультиклассового SVM-классификатора для оценки эффективности. Приводятся результаты, полученные в ходе тестирования на берлинском корпусе эмоциональной речи. В четвертом разделе приводятся оценки эффективности классификации (порядка 85% распознавания на берлинском корпусе для набора из 20 информативных признаков), кратко перечисляются наиболее эффективные информативные признаки и делаются предположения о возможных способах дальнейшего улучшения созданного классификатора.

Ключевые слова: классификация эмоций по голосу, информативные признаки, выбор информативных признаков

1. Обзор информативных признаков

Важнейшим звеном системы автоматического детектирования эмоций по голосу диктора является выделение оптимального набора информативных признаков, коррелированных с эмоциональными состояниями. Выбор информативных признаков оказывает значительное влияние на эффективность классификации. Условно характеристики речи диктора можно разбить на два основных класса — акустические и лингвистические. В зависимости от решаемой задачи их относительная эффективность может быть различной. В модельных эмоциональных базах, в которых речь дикторов соответствует заранее определенному

сценарию, либо в многоязычных корпусах, на первый план выходят акустические параметры, в то время как при работе со спонтанной речью роль лингвистических признаков может оказаться весьма существенной [Schuller *et al.*, 2011].

Из-за нестационарности речевых сигналов во времени для определения их акустических характеристик записи обычно приходится разбивать на небольшие фрагменты, именуемые фреймами. Предполагается, что в их пределах исходный речевой сигнал является квазистационарным. Характеристики речевого сигнала, определенные для каждого фрейма, называются локальными. Однако возможна работа и на уровне интегральных характеристик сигнала. Они определяются путем приложения некоторых статистических функционалов ко всем параметрам, выделенным из сигнала.

Однозначного ответа на вопрос, что лучше подходит для нужд распознавания — локальные или интегральные характеристики, пока нет. Предполагается, что применение интегральных параметров позволяет добиться более высокой точности и скорости классификации. Интегральные характеристики речевого сигнала наиболее эффективны для распознавания состояний, соответствующих различным уровням активации психики. Если же эмоциональные состояния соответствуют примерно одному уровню активности, различаясь, например, по валентности (гнев и радость), то их идентификация посредством анализа интегральных характеристик речевого сигнала окажется затруднена. Кроме этого, недостатками применения интегральных параметров является потеря информации об изменениях речевого сигнала во времени и невозможность применения сложных методов классификации из-за нехватки обучающих векторов. Использование локальных характеристик сигнала позволяет обойти большинство из вышеперечисленных трудностей.

Ранее при разработке систем детектирования эмоционального состояния диктора использовались в основном небольшие наборы информативных признаков (несколько десятков). Однако с течением времени, число выделяемых из звукового сигнала характеристик значительно возросло. Создаваемые на основе опыта экспертов признаки по-прежнему играют важную роль при разработке систем детектирования эмоций, однако все больше внимания уделяется методу «грубой силы» (*brute force extracting*). При этом число выделяемых информативных признаков достигает нескольких сотен, после чего к ним может быть дополнительно применен иерархичный набор функционалов. При таком подходе важнейшее значение приобретают последующие процедуры выбора оптимального набора характеристик, требуемых для эффективного обучения классификатора.

Акустические характеристики голоса могут быть условно разделены на пять категории [El Ayadi *et al.*, 2011]: просодические (частота основного тона, темп речи и т.д.), динамические (фонетическая функция [А. С. Рылов, 2003]), фонационные (отношение гармоник основного тона к шуму, джиттер, шиммер [M. Farrus *et al.*, 2008], …), спектральные (линейные спектральные частоты, кепстральные коэффициенты линейной шкалы частот, кепстральные коэффициенты мел-шкалы частот, …) и энергетические (отношение мощностей в спектральных полосах, оценка мощности сигнала и другие как правило, основанные на энергетическом операторе Тигера). Каждая группа показателей

предназначена для описания отдельных аспектов голоса, и находит свое применение в распознавании эмоциональных состояний.

Выбор эффективных информативных признаков для разделения эмоциональных состояний диктора в значительной степени определяется перечнем детектируемых классов эмоциональной речи: характеристики, основанные на энергетическом операторе Тигера, лучше всего приспособлены для детектирования состояния стресса; для разделения состояний с высоким и низким уровнем возбуждения лучше всего подходят просодические характеристики, такие как частота основного тона и энергия сигнала. Для различения нескольких эмоциональных классов необходимо задействовать спектральные параметры, которые также можно скомбинировать с просодическими характеристиками сигнала. Более точно наиболее эффективный набор информативных признаков может быть определен только на этапе экспериментального исследования.

Перед применением функционалов характеристики нижнего уровня можно подвергнуть фильтрации, трансформациям, можно вычислить их первые и вторые производные. Затем можно применять статистические функционалы, методы аппроксимации кривых, или основанные на специфике человеческого восприятия трансформации. Наиболее популярные статистические функционалы это четыре первых момента (среднее, стандартное отклонение, асимметрия и эксцесс), порядковая статистика (экстремумы и темпоральная информация о них), квартили, диапазоны амплитуд, скорость пересечения нуля, подъем/спад, начало/конец и анализ высшего порядка [Schuller *et al.*, 2011]. При помощи методов аппроксимации (как правило, линейной) кривых можно получить коэффициенты регрессии, например, коэффициент наклона и ошибки линейной регрессии.

На основе проведенного анализа существующих подходов классификации психоэмоционального состояния диктора по его речи были исследованы следующие базовые информативные признаки речи:

- просодические: частота основного тона, темп речи;
- динамические: фонетическая функция;
- фонационные: отношение гармоник основного тона к шуму, джиттер, шиммер;
- спектральные: линейные спектральные частоты, кепстральные коэффициенты линейной шкалы частот, кепстральные коэффициенты мел-шкалы частот;
- энергетические: отношение мощностей в спектральных полосах, оценка мощности сигнала.

С применением к ним следующих функционалов от базовых признаков: статистики высших порядков, первая и вторая производные, поведение кратковременных квантилей, энергетический оператор Тигера.

2. Выбор наиболее информативных и помехоустойчивых параметров речи

Число всевозможных информативных признаков, выделяемых из звукового сигнала, может достигать нескольких тысяч. Далеко не все из них эффективны

для решения задач распознавания эмоционального состояния, а немалая часть потенциально полезных характеристик оказывается избыточными. Перед построением и обучением классификаторов необходимо провести предварительную процедуру отбора информативных признаков. Конечной целью данного этапа работы является выделение релевантного набора характеристик звукового сигнала и декорреляция пространства информативных признаков.

В ранних работах информативные признаки выбирались эвристическими методами, с опорой на опыт экспертов. С течением времени появилась возможность задействовать для этих целей возросшую вычислительную мощность используемого аппаратного обеспечения.

Алгоритмы отбора информативных признаков, чаще всего используемые при разработке систем распознавания эмоционального состояния по голосу, можно условно разделить на два класса — «обертки» (*wrappers*) и «фильтры» (*filters*).

Отбор признаков с использованием «обертки» (*wrapper based selection*) использует оценку работы классификатора в качестве критерия оптимизации внутри замкнутого цикла. К сожалению, даже для сравнительно небольших объемов данных исчерпывающий перебор информативных признаков неприемлем. Необходимо выбрать более приемлемую с точки зрения вычислительной сложности, а значит и более ограниченную и менее оптимальную процедуру перебора. Вероятно, на данный момент наиболее популярной является стратегия линейного последовательного поиска (*sequential forward search)* [Pudil *et al.*, 1994]. Алгоритм начинает с пустого множества и последовательно добавляет к нему наилучшие информативные признаки.

Помимо «оберток», можно использовать «фильтры», к примеру, основанные на методах теории информации либо корреляционном анализе [Hall, 1998]. При этом критерием оптимизации является некоторая функция, соотносящаяся с корреляциями между информативными признаками, приростом информации при их добавлении к набору, определенными метриками в пространстве признаков, статистиками и т. п. Плюсом такого подхода является пониженная, по сравнению с «оберткой», вычислительная сложность алгоритма. В то же время, эффективность работы «фильтра», как правило, ниже, чем «обертки», так как предположения, положенные в его основу, будучи целесообразными в одних случаях, могут существенно нарушаться в других.

3. Эксперимент

3.1. Работа алгоритма автоматического выбора оптимального набора информативных признаков

В результате анализа литературы и предварительно проведенных экспериментов было принято решение использовать в процедуре выбора оптимального набора информативных признаков алгоритм sequential forward feature

selection (SFFS). Суть его работы заключается в том, что на каждой итерации к набору добавляется признак, обеспечивающий наилучшую (для данной итерации) эффективность распознавания.

Мультиклассовый SVM-классификатор строится с использованием библиотеки libSVM[1], в которой данная опция реализована посредством набора классификаторов one-vs-one с последующим голосованием. Это позволяет сразу выбрать наиболее оптимальный набор информативных признаков именно для мультиклассовой классификации, не занимаясь подбором параметров для отдельных классов.

В SVM-классификаторе используется RBF-ядро, а в качестве информативных признаков используются расстояния (интеграл модуля разности маргинальных функций распределения) между функциями распределения различных видов наблюдений и соответствующих медианных функций распределения для различных эмоциональных классов.

После подбора очередного информативного признака параметры модели, используемой при обучении классификатора, корректируются для достижения оптимальной эффективности распознавания.

3.2. Оценка эффективности классификации

Эффективность распознавания на каждой итерации и в процедуре подбора параметров модели оценивается при помощи метода v-fold cross-validation, реализованного в составе пакета libSVM. Данный метод подразумевает разбиение обучающего множества на v частей, с последующим обучением классификаторов на v-1 части, тестированием на оставшейся части и усреднением полученных результатов.

Итоговая оценка сформированного набора информативных признаков определяется как медианное значение эффективности классификации, достигаемой моделью при случайном разделении исходного множества на обучающую и тестовую выборки.

3.3. Тестирование алгоритма SFFS на берлинском корпусе эмоциональной речи[2]

Тестирование алгоритма проводилось на записях, взятых из берлинского корпуса эмоциональной речи (Emo-DB). База включает 535 записей речи 10 дикторов (5 мужчин, 5 женщин), воспроизводящих набор дискретных эмоциональных

---

[1] libSVM — это программный пакет предназначенный для разработки алгоритмов на основе метода опорных векторов;

[2] Берлинский корпус эмоциональной речи — это речевая база, которая содержит записи эмоциональной речи на немецком языке, записанные профессиональными актерами в акустической комнате.

состояний, называемых иногда «архетипическими» (гнев, раздражение, страх, радость, скука, печаль и нейтральное состояние). Авторское исследование Берлинской базы показало [Burkhardt F *et al.*, 2005], что эмоции в ней распознаются слушателями в 80% случаев, и в 60% признаются естественными.

Зависимость эффективности распознавания эмоциональных состояний от числа добавленных в набор алгоритмом SFS информативных признаков показана на рисунке 1.

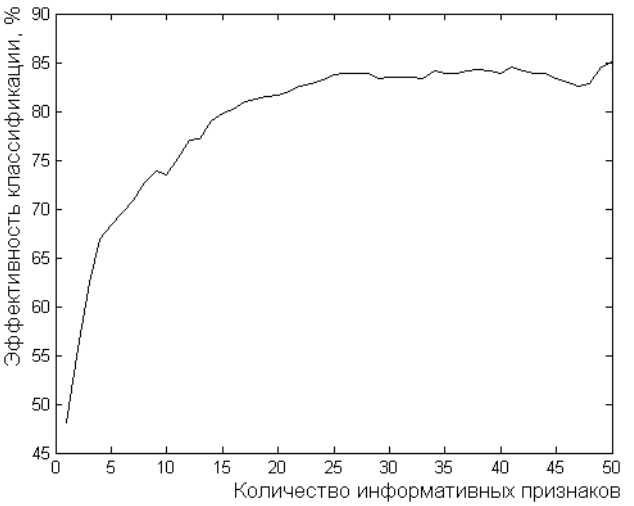


Рис. 1. Эффективность классификации как функция числа добавленных алгоритмом SFS информативных признаков

Видно, что на каждой итерации эффективность распознавания эмоциональных классов возрастала до тех пор, пока не достигла некоторого максимума. Затем эффективность распознавания возрастала с ростом количества признаков очень медленно (Рис. 1).

Тестирование, полученного в результате работы SFS, набора из 50 информативных признаков и подобранных параметров SVM показало, что эффективность классификации составляет приблизительно 85% (матрица спутывания приведена в таблице 1).

Таблица 1. Матрица спутывания классификации эмоциональных состояний

	гнев	**скука**	**раздражение**	**страх**	**радость**	**нейтральное**	**печаль**
гнев	0.8462	0	0	0.0769	0.0769	0	0
скука	0	0.8750	0	0	0	0.1250	0
раздражение	0	0	0.6000	0.4000	0	0	0

	гнев	скука	раздражение	страх	радость	нейтральное	печаль
страх	0	0	0	1	0	0	0
радость	0.1429	0	0	0.1429	0.7143	0	0
нейтральное	0	0.1250	0	0	0	0.8750	0
печаль	0	0	0	0	0	0	1

4. Анализ и выводы

Выбор набора наиболее существенных информативных признаков при помощи sequential feature selection, а так же подбор оптимальных параметров модели SVM-классификатора позволил достичь эффективности распознавания эмоциональных состояний на корпусе Emo-DB порядка 85 %, что сравнимо с таковой для человека.

Для достижения указанной эффективности распознавания, согласно графику, приведенному на рисунке 1, достаточно 20 информативных признаков.

В число наиболее эффективных информативных признаков вошли: производная мощности, первая и вторая производные частоты основного тона, MFCC, LPCC и LSF коэффициенты, отношение мощностей в спектральных полосах, а также некоторые функционалы от указанных параметров.

Дальнейшее совершенствование классификатора представляется авторам в использовании динамических моделей эмоциональной речи.

References

1. *Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W. and Weiss B.* A Database of German Emotional Speech // Proc. Interspeech, 2005.
2. *El Ayadi, M., Kamel, M. S. and Karray, F.*, (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3), pp. 572–587.
3. *Рылов, А. С.* Анализ речи в распознающих системах. — Мн.: Бестпринт, 2003. — 264 с.
4. *Farrus, M., Hernando, J.* (2008) Using Jitter and Shimmer in speaker verification. IET Signal Process., 2009, Vol. 3, Iss. 4, pp. 247–257.
5. *Hall, M. A.* (1998) Correlation-based feature selection for machine learning. Ph. D. Thesis, Hamilton, NZ: Waikato University, Department of Computer Science.
6. *Pudil, P., Novovicova, J., Kittler, J.* (1994) Floating search methods in feature selection. Pattern Recognition Lett. 15, pp. 1119–1125.
7. *Schuller, B., Batliner, A., Steidl, S. and Seppi, D.,* (2011) Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, In Press.

ВЛИЯНИЕ ФАКТОРА СИНТАКСИЧЕСКОЙ РОЛИ АНТЕЦЕДЕНТА НА РАЗРЕШЕНИЕ РЕФЕРЕНЦИАЛЬНОЙ НЕОДНОЗНАЧНОСТИ В РУССКОМ ЯЗЫКЕ[1]

Деликишкина Е. А. (delik.kate@gmail.com),
Федорова О. В. (olga.fedorova@msu.ru)

Московский государственный университет имени М. В. Ломоносова, Москва, Россия

Согласно принятой точке зрения, референциальный выбор осуществляется на основании механизма подсчета активации референта в когнитивной системе. По мнению ряда исследователей, выбор между полной и редуцированной именной группой, осуществляемый говорящим, может быть достоверно предсказан на основании фактора активации. Очевидно, аналогично механизм установления референции устроен и у слушателя: местоимение отождествляется с наиболее активированным в данный момент референтом. Однако ситуация осложняется тем, что в дискурсе может находиться одновременно несколько высоко активированных референтов, и случается так, что ни контекст, ни лексико-грамматическая информация не помогают в разрешении неоднозначности. Пытаясь спрогнозировать референциальный выбор адресата в подобных ситуациях, лингвисты выдвигали различные факторы. В данной работе показано, что носители русского языка более склонны отождествлять неоднозначные местоимения с грамматическими подлежащими, чем с дополнениями. Результаты эксперимента на русском материале согласуются с данными, полученными для других языков (в частности, английского и финского), и свидетельствуют в пользу универсальности фактора синтаксической роли при разрешении референциальной неоднозначности.

Ключевые слова: референция, неоднозначность, анафора, местоимение, подлежащее

[1] Работа выполнена при финансовой поддержке гранта РФФИ №12-06-00268.

1. Введение: референциальная неоднозначность и факторы, влияющие на ее разрешение

Порождая высказывания, говорящий, как правило, заботится о том, чтобы быть понятым адресатом. Поэтому наивному носителю языка может показаться удивительным, что исследователи усматривают потенциальную неоднозначность на каждом шагу. Большое количество лингвистических работ последних лет посвящено изучению неоднозначности и способов ее разрешения в лексике, синтаксисе, семантике и прагматике, а также коммуникативной функции неоднозначности в целом (свежий обзор см. в [Piantadosi et al. 2012]). Исследователей дискурса в первую очередь интересует такое явление, как референциальная неоднозначность[2], то есть наличие у анафора нескольких конкурирующих антецедентов.

Согласно свидетельству ряда исследований (см., в частности, [Kibrik 2011]), выбор между полным и редуцированным референциальным выражением осуществляется на основании активированности референта в рабочей памяти говорящего[3]: местоимения и анафорические нули используются только для отсылки к референтам, имеющим в данный момент высокий уровень активации, в то время как слабо активированные референты называются при помощи полных ИГ. Аналогичным образом механизм референции, очевидно, осуществляется и в процессе обработки речевых сообщений при понимании: адресат отождествляет местоимение с наиболее активированным в его рабочей памяти референтом. Очень часто, однако, сразу несколько высоко активированных референтов претендуют на роль антецедента местоимения. В большинстве случаев эта неоднозначность снимается за счет контекста или различных лексико-грамматических средств (подробнее о классификации средств снятия потенциальной референциальной неоднозначности см. в [Федорова и Успенская 2011]).

Тем не менее, иногда адресат сталкивается с глобальной референциальной неоднозначностью, когда несколько присутствующих в дискурсе референтов с равным успехом могут быть выбраны в качестве антецедента анафорического местоимения. Примером такого контекста является пара предложений:

(1) *Tony Blair shook hands with George Bush in the White House. **He** wanted to discuss the situation in Iraq.*[4]

---

[2] В работах А. А. Кибрика используется синонимичный термин «референциальный конфликт».

[3] Отметим, что ряд исследователей описывает процесс референциального выбора, используя термин «фокус внимания». Мы, вслед за [Kibrik 2011], принимаем модель, согласно которой выбор референциального средства когнитивно связан с уровнем активации референта в памяти.

[4] Пример цитируется по [Järvikivi et al. 2005].

Здесь оба упоминающихся в первом предложении референта — 'Tony Blair' и 'George Bush' (в этом и последующих примерах они выделены подчеркиванием) — могут выступать в качестве антецедента местоимения *he* (выделено полужирным шрифтом), стоящего в начале второго предложения.

Вопрос о том, на каком основании адресат осуществляет выбор в подобных случаях, является предметом активного исследования с 1980-х годов. Лингвистами выдвигались различные гипотезы относительно того, какие факторы оказывают влияние на разрешение анафоры в контекстах с референциальной неоднозначностью. В частности, в работе [Gernsbacher and Hargreaves 1988] постулируется Принцип приоритета первого упоминания (англ. first-mention account). Согласно этому принципу, при выборе антецедента предпочтение отдается тому референту, который был упомянут в предыдущей клаузе первым, вне зависимости от других факторов. Действие этого принципа объясняется общим устройством когнитивной системы человека. В своей аргументации авторы опираются на концепцию построения ментальных структур (англ. structure building framework): для построения связной ментальной структуры необходимо, чтобы сначала была заложена основа, некий стержень, на который будет нанизываться поступающая в дальнейшем информация. Поэтому всякая информация, фигурирующая в начале, является более доступной.

Альтернативная гипотеза изложена в работах [Frederiksen 1981] и [Crawley et al. 1990]. Авторы данных исследований приводят аргументы в пользу Принципа приоритета подлежащего (subject-preference account), согласно которому при осуществлении кореференции предпочтение отдается референту, который был подлежащим предыдущей клаузы. Таким образом, по мнению авторов, разрешение референциальной неоднозначности оказывается обусловлено чисто лингвистическим фактором, а не общими особенностями устройства когнитивной системы человека[5].

Однако разграничить две конкурирующие гипотезы, опираясь на материал такого языка, как английский, оказывается невозможным. В английском, имеющем фиксированный порядок слов, подлежащее почти всегда оказывается упомянутым первым. Таким образом, и Принцип приоритета первого упоминания, и Принцип приоритета подлежащего предсказывают, что в качестве антецедента местоимения в примере (1) с большой вероятностью будет выбран Тони Блэр. Чтобы выяснить, какой из двух факторов в действительности имеет большее значение, авторы работы [Järvikivi et al. 2005] использовали материал финского языка, имеющего свободный порядок слов и кодирующего синтаксические роли с помощью падежного маркирования. Исследователи сконструировали ряд экспериментальных контекстов, состоящих из двух предложений. Первое предложение вводило двух персонажей и могло иметь порядок слов SVO (пример 2) либо OVS (пример 3). Второе предложение начиналось с местоимения, которое могло отсылать к любому из этих персонажей.

[5] Ср., однако, мнение о связи понятия подлежащего с понятием фокуса внимания ([Myachykov and Tomlin 2008]).

(2) *Tony Blair kätteli George Bushia valkoisessa talossa. **Hän** halusi keskustella Irakin tilanteesta.*
[Тони Блэр пожал руку Джорджу Бушу в Белом доме. Он хотел обсудить ситуацию в Ираке.]

(3) *George Bushia kätteli* Tony Blair *valkoisessa talossa. **Hän** halusi keskustella Irakin tilanteesta.*
[Джорджу Бушу пожал руку Тони Блэр в Белом доме. Он хотел обсудить ситуацию в Ираке.]

Эксперимент проводился с помощью методики регистрации движения глаз. Испытуемых просили слушать истории и одновременно с этим смотреть на экран, на котором предъявлялись изображения двух персонажей, в то время как движения их глаз регистрировались с помощью специальной аппаратуры. Статистический анализ полученных данных подтвердил наличие обоих эффектов. С одной стороны, испытуемые, услышав местоимение, чаще фиксировали взгляд на изображении персонажа, выступавшего в первом предложении в роли подлежащего, что свидетельствует о том, что местоимение интерпретировалось ими скорее как кореферентное подлежащему, нежели дополнению. С другой стороны, фиксаций было значительно больше на изображениях персонажа, упомянутого первым, чем вторым, а значит, эффект приоритета первого упоминания тоже имеет место. При этом корреляции между двумя эффектами обнаружено не было, что говорит об их независимости друг от друга. Таким образом, авторы приходят к выводу, что гипотезу, высказанную в [Gernsbacher and Hargreaves 1988], о том, что разрешение референциальной неоднозначности происходит за счет общих когнитивных процессов и непосредственно не связано с лингвистическими факторами, по всей видимости, следует отвергнуть[6]. В заключении авторы подчеркивают, что процесс разрешения анафоры осуществляется под влиянием одновременно нескольких факторов, поэтому от однофакторных моделей референциального выбора следует отказаться[7].

2. Эксперимент М. Г. Степиной на русском материале

Первый подобный эксперимент на русском материале был проведен в 2005 году ([Степина 2005]). Стимульный материал конструировался по тому же принципу, что и описанный в работе [Järvikivi et al. 2005]: экспериментальный контекст состоял из двух предложений, в первом из которых вводились два одушевленных референта одного пола, а второе начиналось с анафорического местоимения *он/она*, которое по смыслу могло относиться к любому из этих двух референтов. Первое предложение имело два варианта — SVO и OVS.

[6] См., однако ссылку 5.

[7] См. аналогичный многофакторный подход в [Kibrik 2011].

Принципиальное отличие от эксперимента на финском материале было в том, что порядок упоминания референтов сохранялся, изменялось лишь их падежное маркирование, для чего дополнительно изменялась диатеза глагола. Это было сделано с целью сохранить тема-рематическую структуру предложения. Так, в примере (4) упоминаемая первой ‘пожилая нянечка’ выступает в функции подлежащего, а ‘новая медсестра’ является дополнением, выражаемым творительным падежом; в примере (5) ‘нянечка’ понижается до статуса прямого дополнения и маркируется винительным падежом, в то время как ‘медсестра’ получает статус подлежащего.

(4) *Пожилая нянечка нашей поликлиники уже давно возмущалась новой медсестрой. **Она** всячески выказывала свое недовольство.*

(5) *Пожилую нянечку нашей поликлиники уже давно возмущала новая медсестра. **Она** всячески выказывала свое недовольство.*

В данном исследовании, проведенном с 36 испытуемыми, использовалась методика чтения с саморегуляцией скорости. Тексты предъявлялись по одному на экране компьютера. После каждого текста появлялся слайд с вопросом, на который предлагалось на выбор два варианта ответа. Испытуемый должен был читать про себя предложение, нажимать на «пробел» и отвечать на вопрос, после чего заносить ответ в специальный бланк. 16 экспериментальных текстов в квазислучайном порядке чередовались с 36 текстами-филлерами, призванными отвлечь испытуемого от истинной цели исследования. Все вопросы к экспериментальным текстам запрашивали референта (*Кто выказывал недовольство?*), вопросы к филлерам относились также к какой-либо фактической детали в тексте.

В результате эксперимента получилось, что на долю первого упоминания приходится 47% всех ответов, на долю второго — соответственно, 53%. Как отмечает автор, «мы наблюдаем небольшую тенденцию к предпочтению выбора последнего из введенных референтов в качестве антецедента неоднозначного местоимения, но расхождение в 6% не может рассматриваться как статистически значимое» [Степина 2005: 48]. Процентное же отношение референтов-подлежащих к референтам-дополнениям оказалось равно 53,6% против 46,4%, что, по словам автора, тоже нельзя считать статистически значимой разницей. Таким образом, проанализировав полученные результаты на предмет соответствия гипотезам о приоритете первого упоминания и о приоритете подлежащего, автор исследования не обнаружила эффектов, выявленных в эксперименте [Järvikivi et al. 2005] для финского языка.

Объясняя полученные результаты, Степина отмечает, что «в эксперименте встретилось несколько примеров, которые оказались нечувствительны к видоизменению структуры» [Степина 2005: 52]: в 4 из 16 экспериментальных контекстов местоимение как при порядке SVO, так и при порядке OVS трактовалось в пользу определенного референта. Так, в приведенных выше примерах (4) и (5) большинство испытуемых посчитало, что антецедентом местоимения

являлась ‘нянечка’. Очевидно, семантический «перевес» в пользу одного из референтов в данном случае настолько силен, что фактор варьирования синтаксической структуры оказался нерелевантным.

3. Новая серия экспериментов на русском материале

Учитывая семантическую несбалансированность как минимум четверти из используемых в исследовании [Степина 2005] экспериментальных контекстов, можно предположить, что отсутствие ожидаемых эффектов объясняется не специфичностью русского языкового материала, а низкой экологической валидностью эксперимента. Кроме того, несмотря на примерно равное распределение случаев выбора в пользу первого/второго референта, а также подлежащего/дополнения, отсутствие формальных статистических подсчетов делает результаты не вполне надежными. Наконец, в работе [Степина 2005] была допущена грубая методологическая ошибка: в отличие от примеров на финском языке, русские примеры (4) и (5) не различаются с точки зрения порядка упоминания референтов, вследствие чего собранный материал не может быть использован для тестирования Принципа приоритета первого упоминания.

3.1. Создание стимульного материала

Чтобы получить более валидные свидетельства за или против Принципа приоритета подлежащего, мы воспроизвели эксперимент Степиной, предварительно проверив стимульный материал на предмет семантической сбалансированности. Для этого мы искусственно создали новые стимульные тексты, местоимения в которых, по нашей интуиции, можно было с равной вероятностью отнести к каждому из двух потенциальных антецедентов, см. (6) и (7).

(6) *Моя младшая сестра очень полюбила свою новую учительницу. **Она** всегда приходила в класс за десять минут до звонка.*

(7) *Моей младшей сестре очень понравилась ее новая учительница. **Она** всегда приходила в класс за десять минут до звонка.*

3.2. Тест на приемлемость

Затем для подтверждения данных интроспекции мы провели тест на приемлемость. Мы предъявляли испытуемым 16 экспериментальных предложений и просили оценить, насколько вероятной является трактовка местоимения в пользу того или иного референта в каждом из них, используя четырехбалльную шкалу Ликерта (от 0 — ‘такое значение не существует для данного местоимения’, до 3 — ‘именно это значение приходит в голову первым, сразу после

прочтения предложения’). 32 испытуемых были распределены так, чтобы каждый читал только один вариант текста. Приемлемыми были признаны тексты, средние оценки которых варьировались от 1,1 до 2,9. Один текст с референтом-дополнением, получившим оценку 0,8, был отбракован и заменен на другой.

Результаты проведенного теста на приемлемость косвенно свидетельствуют в пользу верности гипотезы о приоритете подлежащего: так, из 32 экспериментальных контекстов только в 5 случаях средняя оценка приемлемости референта-дополнения оказывалась выше оценки референта-подлежащего[8].

3.3. Основной эксперимент

После оценки приемлемости созданных нами текстов мы провели основной эксперимент с 32 испытуемыми, распределенным поровну по двум сформированным экспериментальным листам. Помимо 16 экспериментальных текстов в каждый лист было включено 24 филлера. Процедура эксперимента в целом была аналогичной использованной в [Степина 2005]. Принципиальным отличием являлось то, что испытуемому не предлагалось вариантов ответов на поставленные вопросы, чтобы не акцентировать его внимание на потенциальной неоднозначности текстов. Кроме того, испытуемый не записывал ответы, а произносил их вслух.

Количество выбранных в качестве антецедентов подлежащих составило 58,4%, дополнений — 41,6%. Полученные различия являются статистически значимыми (биномиальный тест, $p=0.00017$), а значит, синтаксический статус антецедента следует признать важным фактором при разрешении референциальной неоднозначности в русском языке. Кроме того, общая тенденция к предпочтению подлежащих видна невооруженным глазом: в подавляющем большинстве текстов (в 24 из 32) испытуемые чаще выбирали в качестве антецедента местоимения подлежащее.

Характерно, что 4 из 8 текстов, ответы на вопросы к которым идут вразрез с гипотезой, показали аналогичные результаты в тесте на приемлемость — дополнения в них получили в среднем более высокие оценки, чем подлежащие. Это, по всей видимости, свидетельствует о том, что в данных текстах одна из трактовок имела преимущество перед другой в силу семантического фактора, то есть при дальнейшем усовершенствовании стимульного материала мы можем ожидать еще более показательные результаты.

---

[8] В то же время, общий разброс в оценках приемлемости (от 1,3 до 2,9 для референтов, упоминающихся первыми; от 1,1 до 2,9 для референтов, упоминающихся вторыми) не позволяет сделать аналогичные выводы относительно гипотезы о приоритете первого упоминания.

4. Заключение

Итак, результаты эксперимента подтвердили верность выявленной на материале английского и финского языков гипотезы о влиянии синтаксической структуры предложения на разрешение референциальной неоднозначности и для русского языка. Универсальная тенденция к выбору подлежащего в качестве антецедента максимально активированного референта, выраженного местоимением, свидетельствует, по всей видимости, в пользу того, что когнитивные ограничения накладывают отпечаток на грамматическую структуру языка. Ср., однако, альтернативное мнение из работы [Myachykov and Tomlin 2008], в которой отрицается универсальность Принципа приоритета подлежащего и подчеркивается важность Принципа приоритета первого упоминания[9].

Предполагается, что дальнейшая работа по исследованию референциальной неоднозначности будет вестись по следующим двум направлениям. С одной стороны, для проверки полученных результатов и установления природы когнитивных механизмов, стоящих за обнаруженным эффектом, мы планируем проведение исследований с применением методов регистрации мозговой активности (вызванных потенциалов и функциональной магнитно-резонансной томографии) и движений глаз.

С другой стороны, важной представляется также задача экспериментального тестирования других факторов, способных оказывать влияние на разрешение референциальной неоднозначности[10], а также построения модели, предсказывающей выбор референциального средства на основании этих факторов.

---

[9] Валидность результатов, полученных авторами исследования, с нашей точки зрения, представляется спорной. В качестве стимульного материала эксперимента выступал фильм, в котором красная рыба съедает синюю; при этом внимание испытуемого привлекалось к одной из рыб с помощью стрелки. Задача испытуемого состояла в том, чтобы описать происходящее на экране. Большинство участников англоязычного эксперимента делали рыбу, выделенную стрелкой, подлежащим, меняя активный залог на пассивный в зависимости от того, к какой из рыб было привлечено внимание (Ср. The red fish ate the blue fish; The blue fish was eaten by the red fish). Однако в аналогичном эксперименте на русском материале испытуемые, описывая стимулы, в которых привлекалось внимание к синей рыбе (пациенсу), предпочитали конструкции типа «Синюю рыбу съела красная рыба» конструкциям с пассивом («Синяя рыба была съедена красной рыбой»). Авторы делают вывод: “One initial hypothesis — that there is a universal tendency to assign subject to the perceptually prominent referent regardless of language — is not supported. Indeed, attention to the patient of the event significantly improved its chances of becoming the starting point of a Russian transitive sentence but not its chances of becoming the sentential subject.” ([Myachykov and Tomlin 2008: 44]). Мы считаем, однако, что несоблюдение Принципа приоритета подлежащего в данном случае вероятней всего объясняется нетипичностью для устной русской речи пассивных конструкций, а не нарушением универсальной тенденции к приписыванию наиболее активированному («перцептивно выделенному») референту синтаксической роли субъекта.

[10] В частности, важным представляется исследование на русском материале т. н. Принципа структурного параллелизма, описываемого в ряде зарубежных работ (см., например, [Chambers and Smith 1998]). Согласно этому принципу, предпочтительным

References

1. *Crawley R. A., Stevenson R. J., Kleinman D.* (1990), The Use of Heuristic Strategies in the Interpretation of Pronouns. Journal of Psycholinguistic Research, Vol. 19 (4), pp. 245–64.
2. *Chambers C., Smyth R.* (1998). Structural Parallelism and Discourse Coherence. Journal of Memory and Language, 39, pp. 593–608.
3. *Fedorova O. V., Uspenskaia A. M.* Experimental Analysis of Discourse: The Impact of a Potential Referential Conflict on the Choice of the Referring Expression (On the Material of Russian) [Eksperimental'nyi Analiz Diskursa: Referentsial'nyi Vybor v Situatsii Potentsial'nogo Referentsial'nogo Konflikta (Eksperimental'noie Issledovaniie na Materiale Russkogo Iazyka)]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2011" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011". Bekasovo, 2011, pp. 207–18.
4. *Frederiksen J.* (1981), Understanding Anaphora: Rules Used by Readers in Assigning Pronominal Referents. Discourse Processes, Vol. 4, pp. 323–47.
5. *Gernsbacher M. A., Hargreaves D. J.* (1988), Accessing Sentence Participants: The Advantage of First Mention. Journal of Memory and Language, Vol. 27, pp. 699–717.
6. *Järvikivi J., van Gompel R. P., Hyönä J., Bertram R.* (2005), Ambiguous Pronoun Resolution: Contrasting the First-Mention and Subject-Preference Accounts. Psychological Science, Vol. 16 (4), pp. 260–4.
7. *Kibrik A. A.* (2011), Reference in Discourse. Oxford University Press, Oxford.
8. *Myachykov A., Tomlin R. S.* (2008), Perceptual Priming and Structural Choice in Russian Sentence Production. Journal of Cognitive Science, Vol. 9 (1), pp. 31–48.
9. *Piantadosi S. T., Tily H., Gibson E.* (2012), The Communicative Function of Ambiguity in Language. Cognition, Vol. 122 (3), pp. 280–91.
10. *Stepina M. G.* (2005), The Impact of Working Memory Span on Language Behaviour (on the Example of the Process of Resolving Anaphora) [Vliianie Ob"ema Operativnoi Pamiati Cheloveka na ego Iazykovoie Povedenie (na Primere Vybora Antetsedenta Lichnogo Mestoimeniia)]. Unpublished degree thesis, Moscow State University.

является выбор антецедента, находящегося в той же синтаксической позиции, что и неоднозначное местоимение.

О СИНОНИМИИ ФОКУСИРУЮЩИХ ЧАСТИЦ (НА МАТЕРИАЛЕ НЕМЕЦКОГО И РУССКОГО ЯЗЫКОВ)[1]

Добровольский Д. О. (dm-dbrv@yandex.ru),
Левонтина И. Б. (irina.levontina@mail.ru)

Институт русского языка им. В. В. Виноградова РАН,
Москва, Россия

Как уже многократно отмечалось, немецкий и русский язык отличаются необычайным разнообразием частиц разных типов. В данной работе рассмотрены немецкие частицы *eben* и *gerade* и русские *именно* и *как раз* в фокусирующем значении. В частности, обращается внимание на то, что *gerade* и *как раз* не могут использоваться в качестве ответной реплики, а для *eben* и *именно* такое употребление, напротив, очень характерно. Показано, что блокирование возможности самостоятельного употребления — следствие идеи пересечения. *Gerade* и *как раз* маркируют некое пересечение двух ситуаций, двух представлений и т. п. Однако в случае *eben* и *именно* возможность самостоятельного употребления проистекает из разных семантических источников. Если у *именно* это идея подтверждения, то у *eben* это идея определения самого важного, решающего элемента ситуации.

Ключевые слова: семантика, синонимия, фокусирующие частицы, русский язык, немецкий язык, межъязыковая эквивалентность

1. Постановка задачи

Как уже многократно отмечалось, в том числе и авторами этих строк (см. [Добровольский, Левонтина 2010]), немецкий и русский языки отличаются необычайным разнообразием частиц разных типов. При этом зачастую может создаться впечатление, что между теми или иными русскими и немецкими частицами есть полная эквивалентность. Однако при более внимательном рассмотрении оказывается, что речь, как правило, идет об определенном типе контекстных совпадений, за которыми обычно стоят разные семантические механизмы.

[1] Работа выполнена при финансовой поддержке Программы фундаментальных исследований отделения историко-филологических наук РАН «Язык и литература в контексте культурной динамики» и Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика», грантов РГНФ №№ 10-04-00273а и 11-04-00105а и гранта НШ-6577.2012.6 для поддержки научных исследований, проводимых ведущими научными школами.

В данной работе мы рассмотрим немецкие частицы *eben* и *gerade* в их фокусирующем значении и сравним их с на первый взгляд очень похожими русскими выражениями *именно* и *как раз*. В известном смысле наша работа продолжает исследовательскую традицию, восходящую к монографии [König 1991]. Разумеется, все эти слова очень многозначны.[2] Нас они в данном случае интересуют только в указанном значении. Ср. *именно <как раз> сегодня; eben <gerade> heute*. В большом числе контекстов частицы *именно* и *как раз*, а также *eben* и *gerade* оказываются взаимозаменимыми. При этом в каждой из этих пар лишь одна единица обладает способностью употребляться в качестве самостоятельного высказывания. В немецком языке это частица *eben*, а в русском — *именно*. Частица же *gerade* и фразема *как раз* таким образом употребляться не могут. Ср.

*— Er sollte die Arbeit schon gestern abgeben. — Eben. <*Gerade.>*

*— Он должен был сдать работу еще вчера. — Именно. <*Как раз.>*

Цель данной работы состоит в том, чтобы описать семантику этих четырех единиц, и в том числе обнаружить те семантические особенности, которые стоят за указанным различием.

2. Семантика частицы *eben* и ее русские соответствия

Анализ употребления немецкой частицы *eben* в корпусах[3] позволяет сформулировать следующую гипотезу. В основе семантики этой частицы лежит идея попадания в самую суть, выделения **наиболее существенного** компонента ситуации. Ср.

„Ja, aber wenn ich Ratgeber sein soll, müßte ich wissen, worum es sich handelt", sagte Fräulein Bürstner. „Das ist ***eben*** *der Haken", sagte K., „das weiß ich selbst nicht".* [F. Kafka. Der Prozeß]

---

[2] Так, *eben* употребляется, среди прочего, как наречие времени, близкое по значению к русским *сейчас, сию минуту, только что*, как модальная частица (в значениях, близких к *короче говоря, просто*, с одной стороны, и к *в общем, ведь* — с другой). А *gerade* в своих адвербиальных употреблениях может выражать идею одновременности (ср. *в данный момент*), идею предшествования (ср. *только что*), сочетается с выражениями с пространственным значением (ср. *er wohnt gerade um die Ecke = он живет рядом, за углом*), имеет значения, близкие к *тем более* и к *только-только*. Далее, *gerade* может употребляться как модальная частица (ср. *das fehlte gerade noch! = этого еще только не хватало!*). В своих первичных адъективно-пространственных значениях *eben* и *gerade* значат ‘гладкий, ровный, плоский’ и ‘прямой’ соответственно. См. подробнее DUW и НБНС.

[3] Здесь использован в основном немецко-русский подкорпус параллельных текстов Национального корпуса русского языка (НКРЯ).

*— Да, но, если мне стать вашим советчиком, я должна знать, о чем идет речь, — сказала фройляйн Бюрстнер. — В этом-**то** и загвоздка, — сказал К., — я сам ничего не знаю!* [Ф. Кафка. Процесс. Пер.: Р. Райт-Ковалева]

В переводе данного примера вполне можно было бы использовать и русское *именно*. Тем не менее, определенное семантическое различие между *eben* и *именно* в этом случае сохраняется. Если *именно* подчеркивает, что мнение говорящего полностью совпадает с мнением, высказанным собеседником, то *eben* выражает несколько другую мысль. В случае *eben* говорящий указывает на то, что в реплике собеседника была отмечена самая существенная составляющая ситуации, решающая для ее правильного понимания. По этой причине *eben* невозможно перевести как *именно* в следующем примере, где говорящий реагирует не на мнение собеседника, а на его намерение.

*„Meine Pflicht ist es, dich zu verhaften", erklärte der Kommandant. „**Eben**."* [F. Dürrenmatt. Justiz]

*— Мой долг велит мне арестовать тебя. — **Вот и действуй**.* [Ф. Дюрренматт. Правосудие. Пер.: В. Герасимов]

Очень характерно использование *eben* в тавтологической конструкции, в которой были бы невозможны ни русские *именно* и *как раз*, ни немецкое *gerade*:

*Das Buch ist **eben** das Buch.* [http://log.netbib.de/archives/2003/12/31/das-buch-ist-eben-das-buch/]

Тавтологические конструкции вообще замечательны тем, что в них всегда профилируется то, что является самым важным в значении данного слова (см. [Шмелев 1990]). По этой причине немецкое *eben* очень естественно употребляется в такой конструкции.

Наша интуиция подтверждается тем, что в литературных переводах немецких текстов *eben* часто опускается, поскольку на первом плане в этих контекстах тимиологическая оценка[4], а не согласие с собеседником. По этой причине *именно* там было бы не уместно.

*Man trifft sich hier (im wesentlichen die gleiche Gesellschaft wie unten, nur **eben** festlich kostümiert) nach Premieren und nach den großen Geschäften, nicht um Dinger zu drehen, sondern um gedrehte Dinger zu feiern.* [F. Dürrenmatt. Justiz]

Здесь встречаются (собственно, то же самое общество, что и этажом ниже, только в парадном обличье) после премьер и после крупных сделок,

[4] То есть, оценка по значимости.

однако не затем, чтобы провернуть дело, а затем, чтобы отпраздновать уже провернутое. [Ф. Дюрренматт. Правосудие. Пер.: В. Герасимов]

В русском переводе теряется важная мысль, что *парадное обличье* здесь — не просто отличительный признак, но признак необходимый и существенный.

Анализ параллельных корпусов показывает, что во многих случаях для перевода *eben* используется вовсе не *именно* или *как раз*, а такие выражения, как *в том-то и дело* и *то-то и оно*. Ср.

*„Nur das", sagte sie bestimmt. „**Eben...**"* [F. Dürrenmatt. Justiz]

*— Да, и больше ничего, — сказала она уверенно. — **В том-то и дело.*** [Ф. Дюрренматт. Правосудие. Пер.: В. Герасимов]

Такие способы перевода особенно характерны для сочетания частиц *ja eben*.

*„Das ist es **ja eben**, Herr K.", sagte Frau Grubach, es war ihr Unglück, dass sie, sobald sie sich nur irgendwie freier fühlte, gleich etwas Ungeschicktes sagte.* [F. Kafka. Der Prozeß]

*— **То-то и оно**, господин К., — сказала фрау Грубах, но, к несчастью, как только она чувствовала себя хоть немного увереннее, она сразу становилась бестактной.* [Ф. Кафка. Процесс. Пер.: Р. Райт-Ковалева]

Чтобы пояснить, в каком смысле мы говорим о тимиологической составляющей в значении немецкой частицы *eben,* обратимся к русской фраземе *то-то и оно*, которая, как уже было сказано, часто выступает в роли переводного эквивалента *eben* в определенном типе контекстов. Рассмотрим следующие примеры:

*Вы знаете, что такое алиби? — Нет, не знаю. — **Вот то-то и оно-то**. Для того чтобы установить наличность преступления, я должен прежде всего установить ваше алиби* [А. Т. Аверченко. Ниночка]

*Но я-то вроде свободна! **То-то и оно,** что — «вроде»…* [Н. Катерли. На два голоса]

*Квартира почти продана… почти… **То-то и оно,** что «почти»! Если б не это «почти», я бы не топтался здесь за бесплатно* [А. Волос. Недвижимость]

Как видно из приведенных примеров, здесь говорящий при помощи *то-то и оно* в первую очередь не подтверждает то, что говорит собеседник, а указывает на важность сказанного или выделяет наиболее важный компонент. Ср. *то-то и оно, что вроде* [собеседница говорит *вроде свободна*, для нее

здесь главное — то что она *свободна*, а говорящий указывает, что *вроде* здесь и есть самое важное].

Более того, сочетание *то-то и оно, что* может вводить не только подтверждение, но и опровержение слов собеседника; ср.:

> *— Они же это делают несознательно. —* ***То-то и оно, что*** *несознательно. Знаешь, в детстве я тоже мечтал стать невидимкой.* [М. Сергеев. Волшебная галоша, или Необыкновенные приключения Вадима Смирнова]

> *— Удалось что-нибудь откопать? — Да* ***то-то и оно, что*** *почти ничего! — Коля огорченно махнул рукой.* [В. Белоусова. Жил на свете рыцарь бедный].

Как мы старались показать, кардинальное отличие немецкой частицы *eben* от русского *именно* состоит как раз в наличии в ее значении тимиологического компонента, которое сближает ее с русским *то-то и оно*. Отсюда становится понятным, почему *eben* свободно употребляется в качестве самостоятельной реплики.[5] Такая реплика призвана выразить тот смысл, что собеседник сказал то, что является решающим для данной ситуации. Может быть, он даже сам не понимал, что это-то и есть самое главное, и говорящий указывает ему на это.

3. Семантика частицы *gerade* и ее русские соответствия

Теперь обратимся в фокусирующей частице *gerade*. Анализ ее употребления в корпусах позволяет установить, что в основе значения этой частицы лежит идея **совпадения** в широком смысле.

> *Übrigens könne man* ***gerade*** *bei solchen Gelegenheiten sehen, wie ungemein ernst die Herren ihren Beruf nehmen und wie sie über Hindernisse, die sie ihrer Natur nach nicht bewältigen können, in große Verzweiflung geraten.* [F. Kafka. Der Prozeß]

> ***Именно*** *в таких обстоятельствах становилось виднее всего, насколько серьезно эти господа относятся к своей профессии и в какое отчаяние их приводят препятствия, непреодолимые по самой своей природе.* [Ф. Кафка. Процесс. Пер.: Р. Райт-Ковалева]

Наличие в фокусируемой части высказывания указательного местоимения *такой* (ср. *gerade bei solchen Gelegenheiten = именно в таких обстоятельствах* в приведенном выше примере) показывает, что описываемая ситуация

[5] Хотя способность слова *eben* употребляться в качестве эквивалента предложения отмечается современными словарями (ср. DUW, НБНС, [Métrich, Faucher 2009]), семантические причины, объясняющие такие употребления, ранее не выделялись.

сопоставляется с некой схемой, с некоторыми априорными представлениями о данном типе ситуаций. Ср. также следующий контекст.

*Muß denn **gerade** am Sonntag aufgeräumt werden?* [F. Kafka. Der Prozeß]

*Неужто **именно** в воскресенье надо делать уборку?* [Ф. Кафка. Процесс. Пер.: Р. Райт-Ковалева]

У говорящего существует представление о том, чем положено и чем не положено заниматься в воскресенье. То, что происходит в данный момент («попытка начать уборку») сопоставляется говорящим с этими представлениями. Иными словами, в данном случае *gerade* маркирует совпадение или несовпадение конкретной ситуации с некой заранее заданной схемой. В других же случаях *gerade* указывает на точку пересечения двух ситуаций. Ср. слеующий пример.

*In seinem Zimmer riß er gleich die Schubladen des Schreibtischs auf, es lag dort alles in großer Ordnung, aber **gerade** die Legitimationspapiere, die er suchte, konnte er in der Aufregung nicht gleich finden.* [F. Kafka. Der Prozeß]

В комнате он тотчас же стал выдвигать ящики стола; там был образцовый порядок, но удостоверение личности, которое он искал, он от волнения никак найти не мог. [Ф. Кафка. Процесс. Пер.: Р. Райт-Ковалева]

Имеются две ситуации: «порядок в ящиках стола» и «поиски удостоверения». Частица *gerade* маркирует конфликт в точке их пересечения.

Из сказанного становится ясно, что невозможность использования *gerade* в качестве самостоятельной реплики семантически вполне обоснованна — именно в силу того, что функция этой частицы состоит в маркировании пересечения, а значит, ее семантика не содержит материала для самостоятельного высказывания.

Отметим, что наши наблюдения над особенностями семантики частиц *eben* и *gerade* отличаются от представленных в литературе результатов (ср. в первую очередь известный словарь [Métrich, Faucher 2009]).

4. Русские *именно* и *как раз*: семантика и особенности употребления

Теперь рассмотрим кратко слова *именно* и *как раз* (*Именно там мы и познакомились; Как раз в этот момент зазвонил телефон*) — наиболее близкие аналоги немецких *eben* и *gerade*. Более подробно они описаны в НОССе. Об этих словах см. также [Пайар 1998а; 1998б].

Именно предполагает, что есть некоторое множество объектов, явлений и т. п. и данный объект или явление выделены из этого множества на основании какого-то существенного для говорящего свойства:

На Покров Петр Петрович хотел очаровать всех до единого своим радушием, да и показать, что ***именно*** *он первое лицо в доме.* [И. Бунин. Суходол]

Да, его сиятельство шел ***именно*** *к ней, потому что глядел прямо на нее в упор и слащаво улыбался.* [А. П. Чехов. Анна на шее]

Во многих случаях *именно* подчеркивает, что говорящий имеет в виду в точности то, что он называет:

Здесь расстрелянный Антипов, покойный муж мой, студентом комнату снимал, ***именно*** *вот эту комнату, где мы с вами сидим.* [Б. Пастернак. Доктор Живаго]

В подобных контекстах *именно* сближается со словами *точно, в точности, ровно*. Среди немецких частиц сходной семантики наиболее подходящей в подобных контекстах представляется *genau*, близкая по значению русским выражениям *точно* и *в точности*. Показательно, что ни *eben*, ни *gerade* не передают соответствущий смысл.

В основе *как раз,* в отличие от *именно,* лежит идея совпадения, часто случайного. Это может быть совпадение двух величин, пространственно-временных координат или других свойств какого-либо события или объекта с желаемыми или нежелательными свойствами, слов собеседника с мыслями говорящего, догадки с истинным положением дел и т. п. В этом смысле выражение *как раз* оказывается похожим на немецкую частицу *gerade*, которая могла бы быть употреблена в переводе следующих контекстов.

Четыре часа ***как раз*** *и хватит.* [Л. Петрушевская. Золотая богиня]

Гладышев даже плюнул в досаде и ***как раз*** *не вовремя, потому что раздались общие аплодисменты.* [В. Войнович. Жизнь и необычайные приключения солдата Ивана Чонкина]

Вот это ***как раз*** *я и чувствую, и ты со своей больной душой пришел точно по адресу: у меня тоже болит душа*. [В. Шукшин. Верую!]

Только что забракованное одним приходилось ***как раз*** *по душе другому.* [В. Ходасевич. Торговля]

При этом неправильно **Что ты* ***как раз*** *сейчас делаешь?* [говорящий просто выделяет определенный момент времени, никакого совпадения нет] при естественном *А мы* ***как раз*** *сейчас говорили о тебе!* [появление слушающего случайно совпало по времени с моментом, когда другие его же и обсуждали].

Именно тоже может описывать ситуацию совпадения, стечения обстоятельств, но и в таком контексте сохраняется отличие *именно* от *как раз. Именно*

представляет совпадение не как игру случая, а скорее как проявление закономерности — принципа, воли рока и т. п.

***Именно** я сказал, что передовой не надо, **именно** я предложил поместить эпиграммы и сам написал одну из них.* [А. Рыбаков. Дети Арбата]

*Вокруг шла нормальная жизнь, — и только ей, **именно** ей не повезло!* [Ф. Незнанский, Э. Тополь. Журналист для Брежнева]

Как раз в таких контекстах было бы неуместно или меняло бы смысл высказывания. При перевеоде последнего контекста наиболее подходящей представляется частица *ausgerechnet* (впрочем, возможно и *gerade*, но никак не *eben*), что еще раз доказывает отсутствие полной эквилентности между *eben* и *именно*.

Поскольку *именно* означает, что данный объект выбран скорее всего не случайно, это слово часто используется в вопросительных предложениях вида *Почему именно X?*

*Почему же это **именно** сегодня вдруг не пить!* [С. Довлатов. Соло на ундервуде]

*Почему один стул? Почему **именно** этот стул?* [И. Ильф и Е. Петров, Двенадцать стульев]

В переводе первого из этих контекстов были бы уместны частицы *gerade* и *ausgerechnet*, но никак не *eben*, поскольку здесь мы имеем дело с конфликтом некоторой ситуации и представлениями говорящего о нормальном ходе вещей, а не с указанием на сущностные характеристики обсуждаемого явления. Во втором случае частицы *gerade* и *ausgerechnet* представляются также более уместными; употребление *eben* здесь возможно, но меняет смысл высказывания, превращая вопрос в риторический.

Для *как раз* употребления в подобных вопросительных предложениях нехарактерны или невозможны; ср. **Почему как раз я?; ??Почему как раз в пятницу?*

Именно и *как раз* различаются, далее, с точки зрения сферы действия. *Именно* воздействует на какую-либо часть высказывания — ту, которая наиболее важна для говорящего.

***Именно** Вяземский научил эпиграмму быть конкретной, зубастой и метить не в бровь, а в глаз.* [В. Ходасевич. «Щастливый Вяземский»]

*Наши израненные души ищут покоя вот **именно** за такими кремовыми шторами.* [М. Булгаков. Белая гвардия]

Чтобы достигнуть большей убедительности, говорящий может употребить *именно* несколько раз в одном высказывании, подчеркнув таким образом

разные важные с его точки зрения части. Ср. *именно здесь и именно сейчас.* Ср. также пародийное использование этой частицы в рассматриваемой функции:

Чтобы ***именно*** *вы,* ***именно*** *мне,* ***именно*** *дали,* ***именно*** *двадцать пять рублей.* [Н. Тэффи. Демоническая Женщина]

Как раз также может иметь такую узкую сферу действия, сближаясь с *именно*:

У нас ***как раз*** *сегодня годовщина свадьбы.* [Л. Петрушевская. Сырая нога, или Встреча друзей]

Однако, в отличие от *именно, как раз* может воздействовать и на высказывание в целом:

Ей казалось, что вечерами я слишком долго играю на кларнете, а тут ***как раз*** *подвернулся один инженер, серьезный человек, мы с ней расстались.* [А. Вампилов. Прощание в июне]

На это ему никто не ответил, а Плечевой подтолкнул в спину к столу, там ***как раз*** *стояла свободная табуретка* . [В. Войнович. Жизнь и необычайные приключения солдата Ивана Чонкина]

Другая важная особенность слова *именно*, отличающая его от *как раз*, состоит в том, что оно связывается в первую очередь с подтверждением ранее существовавшей гипотезы или уже высказанного мнения. Этим *именно* отличается и от своего немецкого квазиэквивалента *eben*. В диалоге *именно* часто подтверждает мнение, высказанное собеседником.

— Нет, вы не ослышались, — учтиво ответил Берлиоз, — ***именно*** *это я и говорил*. [М. Булгаков. Мастер и Маргарита]

— Универмаг на Дубнинской? — Да-да, ***именно*** *так*. [Диалоги «09»]

По этой причине *именно*, особенно часто в сочетании с частицей *вот*[6], используется в качестве самостоятельной реплики, выражающей полное согласие с тем, что сказал собеседник.

— ***Именно, именно,*** *— закричал он, и левый зеленый глаз его, обращенный к Берлиозу, засверкал, — ему там самое место!* [М. Булгаков. Мастер и Маргарита]

[6] Не имеет значения, рассматриваем ли мы здесь *вот именно* как сочетание частиц или как отдельную частицу.

— Живы будем — не помрем! — ***Вот именно****, — усмехнулся женский голос.* [Ф. Незнанский, Э. Тополь. Журналист для Брежнева]

— Решили поработать в заповеднике? — ***Вот именно.*** [С. Довлатов. Заповедник]

Отметим, что в подобных контекстах *именно* и особенно *вот именно* в наибольшей степени сближаются с *eben*. Смысловые различия между русской и немецкой частицами, о которых шла речь выше, тем не менее, сохраняются, хотя и едва ли могут привести к невозможности использования этих квазиэквивалентов при переводе.

Использование *как раз* в функции подтверждения затруднено. Высказывание *Вы как раз правы* не может быть сделано просто для выражения согласия с мнением собеседника. Оно предполагает, что, например, кто-то третий оспорил данное мнение, а говорящий считает его правильным.

Иными словами, для *как раз* характерны контексты опровержения, в которых говорящий утверждает нечто сильно отличающееся от ожидаемого. И в этом смысле *как раз* сближается с немецким *gerade*.

Его цель — вовсе не прекращение тех маленьких взрывов или революций, которыми литература движется, а ***как раз*** *наоборот — сохранение тех условий, в которых такие взрывы могут происходить безостановочно.* [В. Ходасевич. Литература в изгнании]

Машина остановилась не потому, что был смертельно ранен командарм — этого в первую секунду ***как раз*** *не заметили.* [К. Симонов. Солдатами не рождаются]

В связи с этим *как раз* (как и немецкое *gerade*), в отличие от *именно* и *eben*, в диалоге часто используется полемически.

Нет-нет! Там ***как раз*** *все по-другому.* [А. Вампилов. Прощание в июне]

— Кто мне там квартиру даст? — Это ***как раз*** *не проблема, па! — оживился Толик.* [В. Кунин. Трое на шоссе]

Простите, ***как раз*** *наоборот — в один какой-нибудь трагичный день вы упадете оттуда и разобьетесь.* [В. Шукшин. Чередниченко и цирк]

Именно также используется в полемических высказываниях, но говорящий в этом случае не столько опровергает слова собеседника, сколько настаивает на своем тезисе.

— Я, — говорю, — репортер. — Журналист? — Нет, ***именно*** *репортер.* [С. Довлатов. Заповедник]

Однако в сознательной, последовательной «социальности» своего творчества он, в сущности, более следует ***именно*** *нелюбимому Гоголю, нежели любимому своему Пушкину.* [В. Ходасевич. Памяти Гоголя]

При этом говорящий не соглашается часто либо с пресуппозицией высказывания собеседника, либо с чем-то, о чем тот умолчал. Если же говорящий прямо отрицает то, что собеседник непосредственно перед этим утверждал, то использование *именно* затруднено или невозможно; ср. *??Вы ослышались, я говорил именно нечто другое, ??Не так, а именно наоборот*.

5. Выводы

Таким образом, даже беглое рассмотрение двух функционально очень близких пар фокусирующих единиц в немецком и русском языках позволяет продемонстрировать сложные и разнонаправленные семантические отношения между ними. Так, мы обратили внимание, что *gerade* и *как раз* не могут использоваться в качестве ответной реплики, а для *eben* и *именно* такое употребление, напротив, очень характерно. Мы считаем, что блокирование возможности самостоятельного употребления — следствие идеи пересечения. *Gerade* и *как раз* маркируют некое пересечение двух ситуаций, двух представлений и т. п. При этом и оба эти выражения в указанном отношении не вполне одинаковы: в *как раз* гораздо сильнее идея случайного совпадения. Тем не менее, самостоятельное употребление блокировано по общей для них причине. Однако там, где возможность самостоятельного употребления имеется (ср. *eben* и *именно*), она проистекает из разных семантических источников. Если у *именно* это идея подтверждения, то у *eben* это идея определения самого важного, решающего элемента ситуации. При описании дискурсивных слов нам вообще представляется чрезвычайно важным проводить различие между функциональным сходством и семантической близостью. Во многих случаях взаимозаменимость (или способность служить переводными эквивалентами друг друга) ничего не говорит нам о семантическом тождестве двух единиц. Это позволяет уточнить наши представления о специфике синонимических отношений в области дискурсивной лексики.

Литература

1. *Добровольский Д. О., Левонтина И. Б.* Диалогические частицы: русско-немецкие соответствия // Логический анализ языка. Моно-, диа-, полилог в разных языках и культурах / Отв. ред. член-корр. РАН Н. Д. Арутюнова. М.: Индрик, 2010. С. 93–103.
2. *НБНС* — Новый большой немецко-русский словарь. В 3 т.: около 500 000 лексических единиц / под общим руководством Д. О. Добровольского. М., 2008–2010.

3. *НОСС* — Новый объяснительный словарь синонимов русского языка / под общим руководством академика Ю. Д. Апресяна. 2-е изд. М., Вена, 2004.
4. *Пайар Д.* Как раз, или миром правит случай // Дискурсивные слова русского языка. М., 1998а.
5. *Пайар Д.* Именно, или как называть вещи своими именами // Дискурсивные слова русского языка / Ред. К. Киселева, Д. Пайар. М., 1998б.
6. *Шмелев А. Д.* Парадоксы идентификации // Тождество и подобие. Сравнение и идентификация. М., 1990.
7. *DUW* — Duden. Deutsches Universalwörterbuch. Hrsg. von der Dudenredaktion. 3., neu bearb. u. erw. Aufl. Mannheim etc.: Dudenverl., 1996.
8. *König E.* The Meaning of Focus Particles. A Comparative Perspective. London: Routledge, 1991.
9. *Métrich R., Faucher E.* Wörterbuch deutscher Partikeln. Unter Berücksichtigung ihrer französischen Äquivalente. Berlin, New York: de Gruyter, 2009.

СЕНТЕНЦИАЛЬНЫЕ ДОПОЛНЕНИЯ С СОЮЗОМ *ЧТОБЫ*

Добрушина Н. Р. (nina.dobrushina@gmail.com)

Национальный исследовательский университет
Высшая школа экономики, Москва, Россия

Формы ирреальных наклонений, такие как субъюнктив или кондиционалис, в языках мира широко употребляются в дополнительных придаточных предложениях. При этом набор глаголов, способных иметь сентенциальный актант в ирреальном наклонении, значительно варьируется от языка к языку. Цель работы – установить, какие предикаты русского языка допускают сослагательное наклонение в дополнительном придаточном. С помощью Национального корпуса русского языка был составлен список таких предикатов и для каждого определена относительная частотность сослагательного и изъявительного наклонений. Обнаружены два семантических типа сентенциальных актантов с сослагательным наклонением: более частотные актанты целевого типа и относительно редкие актанты эпистемического типа.

Ключевые слова: сентенциальные дополнения, дополнительные придаточные, целевые придаточные, сослагательное наклонение, наклонение, семантические типы предикатов, корпус

1. Введение

Сентенциальные актанты в русском языке могут оформляться изъявительным или сослагательным наклонением. Выбор наклонения определяется предикатом (примеры 1а, 1б) или семантикой конкретного контекста (примеры 2а, 2б):

(1а) *Он утверждает, что эта кошка* ***ловит*** *мышей / *Он утверждает,* ***чтобы*** *эта кошка* ***ловила*** *мышей*[1]

(1б) *Нужно,* ***чтобы*** *эта кошка* ***ловила*** *мышей / *Нужно, что эта кошка* ***ловит*** *мышей*

(2а) *Важно, что эта кошка* ***ловит*** *мышей*

(2б) *Важно,* ***чтобы*** *эта кошка* ***ловила*** *мышей*

[1] Вслед за (Brecht 1985) мы считаем, что союз *чтобы* представляет собой сочетание комплементайзера *что* и показателя сослагательного наклонения *бы*.

Сходное явление широко обсуждается для многих других языков: дополнительные придаточные предложения (complement clauses) могут оформляться как индикативом, так и некоторым ирреальным наклонением (например, в немецком, польском, чешском, сербском, литовском, латышском, греческом, во многих романских языках — см. Rothstein & Thieroff 2010, Farkas 2003, Achard 1998, Siegel 2009).

Задача данной статьи — определить условия использования сослагательного наклонения в сентенциальных актантах в русском языке. Процедура исследования состояла из нескольких этапов.

- На основе выборки примеров употребления союза *чтобы* в Национальном корпусе русского языка (НКРЯ) был составлен первоначальный список предикатов, которые могут иметь сентенциальный актант с союзом *чтобы*.
- Этот список был дополнен предикатами, которые упоминаются в некоторых исследованиях (Шведова 1980, Шелякин 1999).
- Для каждого из 57 предикатов списка в подкорпусе с 1970 года было подсчитано количество примеров с сентенциальным актантом в сослагательном наклонении (с союзом *чтобы*) и в индикативе (с союзом *что*). Для этого создавались запросы типа [*принято* расст. от 1 до 1 *чтобы*] и [*принято* расст. от 1 до 1 *что*]. Для тех предикатов, которые могут использовать соотносительные местоимения (*Он претендует* ***на то, чтобы*** *получить медаль*), проверялись четыре конструкции:

  [*настаивать* | *настоять* расст. от 1 до 1 *чтобы*]
  [*настаивать* | *настоять* расст. от 1 до 2 *тот* расст. от 1 до 1 *чтобы*]
  [*настаивать* | *настоять* расст. от 1 до 1 *что*]
  [*настаивать* | *настоять* расст. от 1 до 2 *тот* расст. от 1 до 1 *что*]

В некоторых случаях проверялись также контексты с большей дистанцией между глаголом и комплементайзером, например, для глагола *заставлять*, который часто имеет дополнение (*Как заставить Федю, чтобы он ходил в школу?*):

В полученных выборках все примеры проверялись вручную на соответствие требуемым условиям. Например, необходимо было исключить контексты с целевыми *чтобы*-придаточными вместо дополнительных или с местоименным *что* вместо союзного.

Однако в тех случаях, когда количество примеров превышало 500, вручную обрабатывались только первые 500. Общее количество примеров с сослагательным наклонением выводилось исходя из процента ошибок, обнаруженного для первых 500 примеров.

Таким образом для каждого предиката были получены данные относительно того, насколько часто при нем встречается сослагательное наклонение сравнительно с изъявительным. Результаты помещены в таблицу 1. Конструкции с соотносительным местоимением при предикате и без него рассматривались наравне (см. обсуждение этих конструкций в Князев 2009). Предикаты организованы в список согласно частотности оформления их сентенциальных актантов сослагательным наклонением по сравнению с индикативом.

Таблица 1. Предикаты с сентенциальным актантом, оформленным сослагательным наклонением: соотношение изъявительного и сослагательного наклонений в корпусе с 1970 года

Предикаты	**Сентенциальные актанты с союзом *чтобы* (проценты и количество примеров)**	**Сентенциальные актанты с союзом *что* (проценты и количество примеров)**
хотеть	100% (3817)	0%
хотеться	100% (1161)	0%
желать	100% (219)	0%
быть желательным	100% (133)	0%
пожелать	100% (65)	0%
предпочитать	100% (48)	0%
(не)дай бог	100% (111)	0%
можно	100% (13)	0%
нельзя	100% (74)	0%
надо	100% (500)	0%
необходимо	100% (312)	0%
нужно	100% (490)	0%
просить	100% (386)	0%
требовать	100% (867)	0%
приказать	100% (36)	0%
возражать	100% (45)	0%
стараться	100% (271)	0%
работать	100% (25)	0%
заставлять	100% (5)	0%
разрешать	100% (7)	0%
следить	100% (577)	0%
стремиться	100% (151)	0%
быть направленным	99% (158)	0,6% (1)
быть заинтересованным/ заинтересовать	99,3% (260)	0,7% (2)
добиваться	93,8% (301)	6,2% (20)
заботиться	92% (119)	8% (10)
проголосовать / голосовать	91% (20)	5% (1)
быть выгодным	87% (20)	13% (3)
претендовать	81% (64)	19% (15)
хватать	68% (36)	32% (17)
мечтать	67% (220)	33% (109)
настаивать	62% (280)	38% (169)
быть важным	46% (597)	54% (710)
быть принятым	41% (20)	59% (29)

Предикаты	Сентенциальные актанты с союзом *чтобы* (проценты и количество примеров)	Сентенциальные актанты с союзом *что* (проценты и количество примеров)
быть достаточным	42% (55)	58% (75)
способствовать	39% (46)	61% (73)
быть равносильным	35% (8)	65% (15)
склоняться	33% (35)	67% (71)
быть готовым	29% (55)	71% (134)
состоять	25% (545)	75% (1597)
ждать	23% (149)	77% (506)
сводиться	22% (88)	78% (270)
заключаться	19% (411)	81% (1778)
привыкать	8% (44)	92% (477)
кричать	11% (55)	89% (460)
предупреждать	11% (114)	89% (920)
передавать	10% (27)	91% (254)
представлять	9% (118)	91% (1162)
задумываться	8% (13)	92% (153)
помнить	7% (181)	93% (2296)
сомневаться	2,8% (33)	97,2% (1130)
убедить	2,5% (7)	97,5% (269)
бояться	2% (59)	98% (2603)
напоминать	1% (19)	99% (2040)
воображать	1% (3)	99% (304)
быть приятным	0,3% (1)	99,2% (334)
надеяться	0% (3)	100% (~3800)

2. Семантические типы сентенциальных актантов с союзом «чтобы»

В результате исследования выяснилось, что сентенциальные актанты с союзом *чтобы* распадаются на два типа. Эти типы были названы *придаточными целевого типа* и *придаточными эпистемического типа*.

2.1. Сентенциальные актанты целевого типа

Большая часть предикатов из тех, что имеются в таблице 1, присоединяют сентенциальные акты целевого типа: *хотеть, предпочитать, стараться, способствовать, надо* и другие. Поскольку эти предикаты весьма многочисленны, их удобно распределить по нескольким семантическим

группам. Предложенное ниже деление не претендует на окончательность и имеет рабочий характер.

- Желание: *желать, желательно, пожелать, предпочитать, хотеть, хотеться, стремиться, быть заинтересованным, (не) дай бог, мечтать, надеяться, быть готовым, выгодно, бояться*
- Манипуляция: *требовать, приказать, просить, заставлять, запрещать, кричать, предупреждать, передавать, напоминать, возражать*
- Деонтическая возможность и необходимость: *надо, необходимо, можно, нельзя, важно*
- Достижение, цель, результат: *стараться, следить, работать на то / над тем, быть направленным на то, голосовать, добиваться, заботиться, претендовать, настаивать, способствовать, склоняться, выгодно, достаточно, хватать*
- Генерические и хабитуальные ситуации: *быть равносильным, быть принятым, привыкать*
- Пустые глаголы (предикаты тождества) и существительные цели: *состоять, сводиться, заключаться, задача, цель*

Дополнительные придаточные предложения с сослагательным наклонением при этих предикатах обнаруживают ряд характерных особенностей.

Во-первых, они обладают свойством иметь зависимую временную референцию (time reference dependency — Noonan 2007: 102). Их время определено матричным предикатом: они относятся к моменту, который следует ситуацией главного предиката. Исключением является глагол *отвыкнуть*, временная референция которого предшествует главному предложению, и глагол привыкнуть, имеющий вневременную ситуацию в качестве сентенциального актанта. Тем самым, эти предикаты тоже имеют фиксированную временную референцию.

Во-вторых, дополнительные придаточные этого типа, как правило, не охарактеризованы по параметру истинности. Говорящий не в состоянии оценить, является ли ситуация, обозначенная этой пропозицией, истинной или ложной (см. об этом также Borščev & al. 2007, Kagan 2007). Этому принципу, однако, не подчиняются два глагола, которые предполагают истинность подчиненных им пропозиций: *добиться* и *заставить*.

Не имея возможности детально рассмотреть все предикаты, входящие в эту группу (см. подробнее Dobrushina 2012), обсудим лишь некоторые свойства контекстов таких предикатов, которые оформляют зависимые предикации как сослагательным наклонением, так и индикативом. Сравнение этих контекстов поможет нам понять, какие семантические компоненты отвечают за выбор наклонения.

Желание (намерение) является одним из компонентов семантики предикатов *быть заинтересованным, мечтать, надеяться, быть готовым, ждать*. Для глагола *надеяться* основным значением является оценка вероятности подчиненной ситуации, поэтому сослагательное наклонение встречается крайне редко (3 примера из приблизительно 3800):

(3) *Нам опротивело настоящее, мы* ***не надеемся, чтобы*** *жизнь можно было направить по доброму пути…* [О. В. Волков. Из воспоминаний старого тенишевца (1988)]

Глагол *ждать* присоединяет клаузу с сослагательным наклонением тогда, когда означает ‘хотеть, чтобы нечто произошло’ (пример (4)), и с изъявительным — когда означает ‘думать, что нечто произойдет’ (пример (5)):

(4) *Переключая в летнюю жару телевизионные каналы, зритель* ***ждет, чтобы*** *его рассмешили.* [«Известия», 2002.08.05]

(5) *Жена встретила его, как обычно, он ни о чём её не спрашивал,* ***ждал, что*** *сама скажет о своём свидании у водоразборной колонки.* [Василь Быков. Бедные люди (1998)]

Для этих глаголов вероятность оформления зависимой клаузы сослагательным наклонением тем меньше, чем сильнее в их семантике эпистемический компонент. Выбор сослагательного наклонения связан с вовлеченностью субъекта главной предикации в ту ситуацию, которая обозначена зависимой клаузой.

Глаголы речевой манипуляции *кричать, предупреждать, напоминать, настаивать* имеют сослагательное наклонение тогда, когда они обозначают каузацию, и индикатив — тогда, когда каузативный компонент значения отсутствует. Заметим, что только в первом случае придаточное обладает свойством зависимой временной референции:

(6) *И полярники потащили свои вещи, а мама стала* ***кричать, чтобы*** *Алешка шел домой одеваться. [А. Ф. Членов. Как Алешка жил на Севере (1978)]*

(7) *Отец распалялся,* ***кричал, что*** *за себя не отвечает, снимает с себя ответственность… [Юрий Трифонов. Дом на набережной (1976)]*

Таким образом, манипулятивные глаголы присоединяют придаточные с сослагательным наклонением тогда, когда они содержат в своем значении каузативный компонент.

Предикаты достижения содержат значение цели: субъектом этих предикатов обычно является человек, который стремится прийти к некоторому результату, или ситуация, которая является причиной наступления другой: ***Следил, чтобы*** *я сидела дома*. Представляет интерес глагол *способствовать*. Сослагательное наклонение используется при этом глаголе только тогда, когда пропозиция не имеет истинностного статуса. Индикатив тоже может использоваться в нефактивных придаточных:

(8) *Как мог, я* ***способствовал тому, чтобы*** *роман вышел в свет. [Г. Я. Бакланов. Жизнь, подаренная дважды (1999)]* / ОК: *Как мог, я* ***способствовал тому, что*** *роман выйдет в свет.*

Выяснилось, что союз *чтобы* значительно более характерен для таких конструкций с глаголом *способствовать*, где субъект главного предложения осознанно воздействует на ситуацию зависимого, в то время как союз *что* — для таких, где связь между двумя ситуациями является ненамеренной:

(9) *И не только ему — многим художникам, чья ложная скромность невольно* ***способствует тому, что*** *бесталанные, но весьма активные в борьбе личности пробиваются в «лидеры»… [Евгений Весник. Дарю, что помню (1997)]*

Прослеживается корреляция между одушевленностью субъекта матричного предиката и выбором союза: при одушевленном субъекте чаще выбирается *чтобы*.

Таблица 2. Зависимость наклонения от одушевленности субъекта в сентенциальных дополнениях при глаголе «способствовать»

	одушевленный субъект	**неодушевленный субъект**
способствовать *тому, что*	5 % (4)	**95 % (70)**
способствовать *тому, чтобы*	**38 % (17)**	**62 % (28)**

Таким образом, выбор сослагательного наклонения часто коррелирует со степенью вовлеченности субъекта главного предложения в зависимую ситуацию.

2.2. Сентенциальные дополнения эпистемического типа

Для всех рассмотренных выше глаголов эпистемическое употребление блокировало возможность использования сослагательного наклонения в придаточном предложении. Есть, однако, особый тип использования союза *чтобы*, когда дополнительное придаточное предложение имеет эпистемическое значение. Такие контексты упоминаются в русской грамматике (Шведова 1980: 477).

Дополнительные придаточные с сослагательным наклонением эпистемического типа встречаются при некоторых предикатах пропозициональной установки и восприятия: *сомневаться, представлять, воображать, верить/ поверить, помнить, припомнить, находить, маловероятно, не думать, не знать, невероятно*. Предикатов этого типа значительно меньше, и частотность использования при них сослагательного наклонение существенно ниже. Этот тип употребления *чтобы* заметно отличается от предыдущего.

Во-первых, такие дополнительные придаточные не являются зависимыми с точки зрения своей временной референции: они могут относиться к прошлому, настоящему или будущему, вне зависимости от временной референции матричного предиката.

Во-вторых, ситуация придаточного предложения имеет отчетливый истинностный статус. Субъект матричного предиката выражает свою оценку

истинности ситуации, обозначенной в придаточном (truth-value dependent clauses — Noonan 2007: 102). Все предикаты этой группы подразумевают отрицание вероятности встроенной ситуации.

Глагол *сомневаться* выделяется тем, что только он может иметь сослагательное наклонение тогда, когда используется без отрицания. Напротив, в положительной форме *сомневаться* сочетается только с индикативом, так как обозначает, что ситуация высоковероятна.

(10a) ***Сомневаюсь, чтобы*** *не выданные Жанне деньги возвратились в казну.* [Евгений Рубин. Пан или пропал. Жизнеописание (1999–2000)]

(10b) *****Не сомневаюсь, чтобы*** *не выданные Жанне деньги возвратились в казну.*

В целом как для отрицательных, так и для положительных вариантов глагола *сомневаться* сослагательное наклонение не является частотным, причем похоже, что частотность такого оформления снижается по сравнению с 19 веком. Обычный способ оформить зависимое придаточное предложение — это индикатив.

Другие предикаты присоединяют сослагательное наклонение под действием ряда факторов. Следующие факторы способствуют использованию сослагательного наклонения:

- отрицание (11c),
- вопросительность (11d)
- редкость или маловероятность ситуации (11e)

(11a) ***Представь, что*** *лежишь на берегу моря где-нибудь в самом тёплом и приятном месте на земле.* [«Даша», № 10, 2004]

(11b) ?? ***Представь, чтобы ты*** *лежал на берегу моря где-нибудь в самом тёплом и приятном месте на земле.*

(11c) *Но* ***не представляю, чтобы*** *Никита Сергеевич мог так глупо и примитивно рассуждать.* [Анастас Микоян. Так было (1971–1974)]

(11d) *Вы* ***можете себе представить, чтобы*** *подобное случилось с Михаэлем?* [«Формула», 2002.06.15]

(11e) ***Трудно себе представить, чтобы*** *власть так легко отказалась от использования СМИ в своих целях.* [«Известия», 2002.09.24]

Этими свойствами обладает ряд ментальных предикатов, предикатов восприятия (см. об этом также Кобозева 1988), а также наречия пониженной частоты или малой вероятности *вероятно / невероятно / маловероятно*.

(12a) *Юра* ***редко*** *видел, чтобы отец ходил с палкой…* [Анна Берсенева. Возраст третьей любви (2005)]

(12b) *Если взглянуть на историю комедийного жанра, то очень* ***редко встретишь, чтобы*** *главным героем веселого произведения было лицо положительное, светлое, доброе.* [Эльдар Рязанов. Подведенные итоги (2000)]

(12c) ***Вероятно ли****, чтобы Европа была населена, а другая часть света нет?* [Владимир Горбачев. Концепции современного естествознания (2003)]

(12d) *Просто* ***невероятно, чтобы*** *можно было так исхудать меньше чем за год.* [Е. С. Гинзбург. Крутой маршрут (1990)]

(12e) ***Маловероятно, чтобы*** *в этой газете кто-нибудь владел древнегреческим языком.* [Вячеслав Пьецух. Русские анекдоты // «Знамя», 2000]

Все эти факторы — отрицание, вопросительность и низкий эпистемический статус подчиненной предикации сигнализируют о том, что истинностный статус ситуации не может быть установлен. Тот же набор факторов влияет на использование сослагательного наклонения в относительных придаточных (Dobrushina 2010).

3. Заключение

Таким образом, случаи употребления комплементайзера *чтобы* распадаются на два хорошо диагностируемых класса.

Первый класс, названный употреблением целевого типа, охватывает множество предикатов. Этот класс включает предикаты желания, манипуляции, деонтической необходимости и возможности, достижения и другие. Их зависимые предикации имеют следующие свойства:

(а) они обозначают ситуацию, осуществление которой связано с субъектом главного предложения (зависит от него или является его желанием или потребностью)

(б) они, как правило, не имеют истинностного статуса: вероятность их осуществления находится вне оценки (за исключением глаголов *добиться* и *заставить*, которые подчиняют реальные клаузы (*добился, чтобы его пустили внутрь*).

(в) они обладают зависимой временной референцией (относятся к моменту, следующему за осуществлением главной ситуации, иногда — перед (*отвыкнуть*)).

Помимо семантических сходств, дополнительные придаточные целевого типа имеют и другие сходства с придаточными цели. Во-первых, в случае совпадения субъектов главного и зависимого предложений, они тоже используют инфинитив:

Он хочет попрощаться

Он готов к тому, чтобы попрощаться
Он пришел, чтобы попрощаться

Во-вторых, многие из них используют местоименное соотносительное слово *то,* как и целевые конструкции:

Я настаиваю (на том), чтобы ты не уезжал
Я сделал это (для того), чтобы ты не уезжал

Я предполагаю, что основной причиной употребления сослагательного наклонения в этих придаточных предложениях является их близость к целевым конструкциям. Сходство целевых придаточных и сентенциальных актантов при глаголах желания и намерения широко обсуждается в лингвистической литературе (Palmer 2001: 131 — "purpose clauses express what the subject wants or intends"). Более того, хорошо известно, что дополнительные придаточные определенного типа часто развиваются в языках из целевых конструкций (Haspelmath 1989, Dixon 2006: 39, Schmidtke-Bode 2009: 176). История развития союза ч*тобы*, по-видимому, не исследована, но союз *дабы*, согласно работе (Молдован 1996), в древнерусских текстах использовался как целевой, а в 18 веке, согласно НКРЯ, он используется и в сентенциальных актантах.

Напротив, употребления ч*тобы* второго, эпистемического типа, ограничены очень небольшим набором предикатов. Эти предикаты выражают низкую эпистемическую оценку ситуации в контексте отрицания, вопроса или наречий с эпистемическим значением. Эти предикаты не подразумевают вовлеченность субъекта главного предложения в ту ситуацию, которая выражена зависимым. Придаточные этого типа не имеют свойства зависимой временной референции:

(а) *Я не думаю, чтобы он это сделал сам*
(б) *Я не думаю, чтобы он был здесь*
(в) *Я не думаю, чтобы через год кто-то сюда вернулся*

С формальной точки зрения эти придаточные предложения тоже отличаются от целевых. Они не принимают инфинитива:

Я сомневаюсь, чтобы я захотел вернуться
**Я сомневаюсь, чтобы вернуться*

Для них нетипично использование местоименного слова *то*, хотя несколько примеров использования глагола *сомневаться* с местоимением и союзом ч*тобы* в корпусе имеется, почти все — в несовременных текстах.

Можно предположить, что дополнительные придаточные предложения целевого типа развились из целевых придаточных, в то время как эпистемические придаточные имеют какое-то другое происхождение. Проверка этой гипотезы требует отдельного исторического исследования.

Литература

1. *Шелякин М. А.* (1999). Об инвариантном значении и функциях сослагательного наклонения в русском языке. Вопросы языкознания. № 4. С. 124–136.
2. *Achard, Michel.* (1998). Representation of Cognitive Structures: Syntax and Semantics of French Sentential Complements (Cognitive Linguistics Research, 11) (Cognitive Linguistic Research). Mouton de Gruyter.
3. *Брехт, Р. Д.* (1985). О взаимосвязи между наклонением и временем: синтаксис частицы бы в русском языке // Новое в зарубежной лингвистике, XV. Современная зарубежная русистика. С. 101–117.
4. *Князев М. Ю.* (2009). Предикатные актанты с местоимением « то » в русском языке: грамматический статус и особенности употребления. Дипломная работа.
5. *Кобозева И. М.* Отрицание в предложениях с предикатами пропозициональной установки // Логический анализ языка. Знание и мнение. М.: Наука. 1988. С. 82–98.
6. *Молдован А. М.* (1996). Из синтаксиса древнерусского перевода Жития Андрея Юродивого // Русистика. Славистика. Индоевропеистика. Сборник к 60-летию А. А. Зализняка. М. С. 256–275.
7. *Шведова Н. Ю.* (ред.) Русская грамматика. 1980. М.: Наука.
8. *Borschev, V., Paducheva E., Partee B., Testelets Y., Yanovich I.* (2007). Russian Genitives, Non-Referentiality, and the Property-Type Hypothesis. In: Antonenko, Andrei et al., eds. Formal Approaches to Slavic Linguistics: The Stony Brook Meeting 2007 (FASL 16) (pp. 48–67). Ann Arbor: Michigan Slavic Publishers.
9. *Dixon, R. M. W.* (2006). Complement clauses and complementation strategies in typological perspective. In: Dixon, R.M.W. & Aikhenvald, Alexandra Y. (eds.) Complementation: a cross-linguistic typology (pp. 1–48). Oxford: Oxford University Press.
10. *Dobrushina Nina.* (2012). Subjunctive complement clauses in Russian. Russian Linguistics. 36(2).
11. *Dobrushina Nina.* Subjunctive in relative clauses. Oslo studies in language, # 2. 181–210.
12. *Farkas, Donka.* (2003). Assertion, Belief and Mood Choice. Paper presented at ESSLLI, Conditional and Unconditional Modality Workshop, Vienna. http://people.ucsc.edu/~farkas/papers/mood.pdf
13. *Haspelmath, Martin.* (1989). From purposive to infinitive — a universal path of grammaticization. Folia Linguistica Historica 10: 287–310.
14. *Kagan, Olga. (2007). On the Semantics of Structural Case: Hebrew University: Ph.D dissertation.*
15. *Noonan, Michael.* (2007). Complementation. In Timothy Shopen (ed.), Language typology and syntactic description (2nd ed.), vol. II, Complex constructions (pp. 52–150). Cambridge: Cambridge University Press.
16. *Palmer, Frank R.* (2001). Mood and Modality, 2nd ed. Cambridge University Press.

17. *Rothstein B., Thieroff R.* Mood in the Languages of Europe. Amsterdam–Philadelphia: John Benjamins Publishing Company. 2010.
18. *Schmidtke-Bode, Karsten.* (2009). A Typology of Purpose Clauses. John Benjamins, Amsterdam / Philadelphia.
19. *Siegel, Laura.* (2009). Mood selection in Romance and Balkan. Lingua 119, 1859–1882.

ФУНКЦИИ ЧАСТИЦ ОБРАТНОЙ СВЯЗИ В ТЕЛЕФОННОМ ДИАЛОГЕ (НА ПРИМЕРЕ ЛЕКСЕМ *А*, *АГА* И *ГМ*)[1]

Герасименко О. А. (olga.gerassimenko@gmail.com)

Тартуский университет, Тарту, Эстония

Работа описывает функции трех частиц обратной связи на основе предшествующего и последующего диалогического контекста. Анализ 45-минутного корпуса регистратурных звонков показывает, что говорящие используют *гм* для указания на контекстуальную неадекватность предыдущей реплики, *а* для указания на необходимость прояснения и *ага* для указания на достаточность предыдущей реплики и завершенность секвенции.

Ключевые слова: обратная связь, реактивные реплики, спонтанный диалог, конверсационный анализ

Введение

В диалогической речи важное место занимают реплики обратной связи (тж. *реактивные реплики, реплики-реакции*) — конвенциональные реакции на завершенную или продолжающуюся реплику собеседника, которые сигнализируют, что сообщение принято (полностью или частично) и понято определенным образом (Allwood et al. 1993, Yngve 1970, Gardner 2001, Sorjonen 2001). Средства обратной связи могут включать лексические повторы, смех или завершения реплики за собеседника (Stubbe 1998), однако чаще всего обратная связь реализуется в виде минимальных реакций: одно- или двусложных лексем, которые традиционная грамматика (напр., РГ 1980 I: § 1689–1690) относит к классу междометий или частиц[2]. Значение этих лексем описывается в словарях при помощи перечисления эмоций, которые они могут выражать в зависимости от ситуативного контекста. Так, диапазон эмоций, выражаемых междометием *а*, по данным «Большого толкового словаря» (2006) включает припоминание, догадку, узнавание при встрече, удовольствие, радость, удивление, досаду, негодование, решимость с оттенком досады, отчаяние. Некоторые из перечисленных эмоций совпадают с описанием в этом же словаре лексемы *ага*, в междометном значении выражающем «злорадство, торжество,

[1] Исследование проведено при поддержке грантов ETF8558, ETF9124 и программы EKT8 «Эстонская языковая технология 2011».

[2] Подробнее об их теоретическом и практическом разграничении см. Златев 2004, Шаронов 2008, Герасименко 2008.

замешательство и т. п.»; а замешательство как возможный компонент значения *ага* оказывается присущим и междометию *гм*, выражающему «сомнение, недоверие, иронию» и употребляющемуся «при затруднении или нерешимости сказать или сделать что-либо». В разных словарях набор эмоций, описывающих значение той или иной лексемы, может варьировать: например, в «Словаре русского языка» (1999) междометие *ага* выражает «догадку, радостное удивление», а междометие *а* в числе прочего «угрозу, злорадство» (ср. с толкованиями БТС). Большинству толковых словарей свойственны незавершенность списка перечисляемых эмоций и повторение одних и тех же эмоций при описании разных лексем. И. А. Шаронов, автор современного масштабного очерка о русских междометиях, отмечает, что это в общем характерно для словарного описания эмоциональных и когнитивных междометий (Шаронов 2008: 125), и указывает на нечеткость эмоциональных номинаций и неоперациональность толкований, построенных на их перечислении (2008: 203). Шаронов предлагает использовать для истолкования не только наименование эмоции, но и тип стимула, на который она реагирует, отмечая особый характер акта понимания/осознания в качестве стимула когнитивных междометий (2008: 272); кроме того, отдельно указывает на функцию обратной связи в диалоге (2008: 250), что дает возможность описывать значение и функционирование лексем значительно более насыщенно.

Однако в аналитической части исследования И. А. Шаронова разные лексемы в ряде контекстов все же получают идентичное истолкование, не проясняющее разницу в их использовании. (Так, «возглас воодушевления при умозаключении», «возглас злорадства при слабости позиции оппонента, соперника» являются общими для междометий *ага* и *а*, 2008: 226–227; «возглас разочарования при оценке полученной информации» несет дополнение «как не оправдавшей ожиданий» в случае *а*, однако неясно, исключен ли этот дополнительный контекст для *ага*). В других исследованиях, рассматривающих интересующие нас лексемы в рамках когнитивных междометий (Wierzbicka 1991, Добрушина 1995), междометий догадки (Иомдин 2004) или частиц-реакций (Rathmayr 1985: 185–216), различия в использовании тех или иных лексем одного класса также рассматриваются не всегда: исследование Б. Л. Иомдина (2004: 241–244) анализирует речевые ситуации, общие для всего класса междометий догадки; в работе Р. Ратмайр (1985: 198–199) употребление частиц-реакций описывается при помощи лаконичных метаязыковых описаний, повторяющихся для разных лексем (напр., *gut, richtig, klar* описывает значение как частицы *угу*, так и частицы *да*); Н. Р. Добрушина, разграничивая контексты частиц *а (ах)* и *ага/угу* как реакции на запрошенную и незапрошенную информацию (1998: 138–139), отмечает, что *угу* и *ага*, очевидно, не тождественны, однако автор не различает их из-за невозможности выявить различия на материале исследования. Таким образом, во многих исследованиях русских частиц обратной связи общие контексты часто рассматриваются с точки зрения общности значения частиц, а не потенциальных различий. Обратным примером являются детальные исследования Е. А. Гришиной (2009, 2010) на базе мультимедийного корпуса НКРЯ, рассматривающие типы вокальных жестов *а* и *о*, их интонационных рисунков

и сопровождающей их телесной жестикуляции. Спектр использования диалогических частиц в сценарной речи, однако, обычно отличается от их использования в спонтанной речи: в сценарной речи частицы используются для расставления акцентов, а не для управления спонтанно текущим диалогом.

Нам представляется, что с точки зрения осознанного использования средств обратной связи в русском диалоге (например, в изучении русского языка как иностранного или в компьютерном моделировании диалога) наибольший интерес представляют именно систематические различия в использовании частиц, которые обуславливают выбор конкретной частицы в спонтанном речевом взаимодействии в конкретный момент времени. Влияет ли выбор разных частиц из класса функционально схожих лексем на дальнейшее течение диалога?

Рассмотрим спонтанные диалогические контексты лексем *а*, *ага* и *гм*, относящихся к классу реакций на поступающую информацию (Шаронов 2008: 251) и характеризующихся общностью стимулов.

Метод и корпус

Сопоставительное исследование контекстов частиц проведено при помощи инструментария и понятийного аппарата конверсационного анализа (Исупова 2002, Liddicoat 2007). Это квалитативный метод исследования спонтанных речевых ситуаций; он основывается на предпосылках линеарной упорядоченности диалогической речи, ее структурной организованности, контекстуальной ориентированности коммуникативных единиц и экспертности участников диалога. Интерпретации высказываний строятся на основе последующих реплик (*next turn proof procedure*, Hutchby, Wooffitt 2008: 13–15): временная соотнесенность реплик и маркеры (не)предпочтительных реакций сигнализируют, насколько коммуникативный вклад одного собеседника соответствует ожиданиям другого, и тем самым напрямую свидетельствуют о смысле, вложенном в предшествующее высказывание самими собеседниками. При этом учитывается секвенциальная и тематическая структура диалога, а также закономерности смены говорящего: паузальность, указывающая на передачу права говорить, и наложение или слияние реплик, указывающее на перехват очередности (Liddicoat 2007: 79–86). Таким образом, конверсационный анализ учитывает не только левый, стимульный, но и правый, интерпретационный, контекст высказывания. Именно в нем отражается истолкование лексемы самим говорящим или его собеседником, а также, возможно, последующая оценка этой интерпретации (эксплицируемая в случаях расхождения). Отклонения от обычной коммуникативной модели (*deviant cases*) обогащают анализ знанием, насколько (иначе) в них проявляются ожидания и интерпретации собеседников. Для обобщения результатов конверсационный анализ, как правило, пользуется так называемой интуитивной статистикой, оперирующей понятиями «частый», «редкий», «единичный» (Schegloff 1993: 99) в силу внимательности к контекстуальности коммуникативных явлений, слабо выразимой в абсолютных числах или процентах.

Исследование проведено на корпусе телефонных разговоров в регистратуре поликлиники (Герасименко 2008). Корпус состоит из 81 диалога и составляет ок. 45 минут звучания (≈ 9000 словоупотреблений). Записи сделаны в 2001–2003 в русскоязычном регионе Эстонии, характеризующемся изолированностью от эстонского языка и тесной языковой и медиа-потребительской близостью с Россией. Тематическая общность диалогов облегчает сопоставление контекстов, а наличие аудиозаписи — проверку и уточнение транскрипции[3]. Кроме того, телефонное общение исключает жестовую и мимическую обратную связь, которая может усложнять картину анализа.

Корпус содержит входящие и исходящие регистратурные звонки. Говорящие анонимизированы в транскрипции при помощи ритмически равнозначной замены всех идентифицирующих данных. Диалоги транскрибированы в традиции конверсационного анализа (Jefferson 2004, см. приложение), уделяющей особое внимание точной временной соотнесенности всех вокализаций, интонационному и ритмическому рисунку реплик и длительности коммуникативных пауз.

В силу ориентированности на восприятие слушающего, а не на абсолютные измерительные величины, Джефферсон различает четыре типа интонационного контура: 1) плоский, немаркируемый, 2) контур незаконченности (незначительное нисхождение или восхождение тона, в транскрипции «,»), 3) контур завершенности (значительное нисхождение тона, «.»), 4) контур вопросительности (значительное восхождение тона, «?»).

Частицы *а*, *ага* и *гм* в спонтанном телефонном диалоге

Регистратурным диалогам присуща четкая структура, обычно хорошо известная собеседникам. Реплики, как правило, коротки, говорящие сменяются плавно (ожидания собеседников друг от друга ясны и редко конфликтуют). Диалог организован как последовательность вопросо-ответных (тж. директиво-ответных) секвенций с разнообразными наращиваниями. Частицы обратной связи обычно встречаются в третьей позиции секвенций, где инициатор секвенции сообщает, что (полный или частичный) ответ принят и понят; вместе с тем он может перехватить инициативу (например, задать новый или уточняющий вопрос, инициировать завершение диалога) или же оставить/передать очередность говорить собеседнику. Все три частицы в корпусе регулярно встречаются и после незапрошенной информации, хотя контексты с запрошенной информацией значительно больше часты в справочных диалогах.

---

[3] К сожалению, подкорпус спонтанной устной речи НКРЯ, значительно более широкий как территориально, так и тематически, доступен лишь в письменной и довольно упрощенной транскрипции, не дающей представления о точном месте произнесения частиц по отношению к реплике-стимулу, как и о других временных и интонационных подробностях.

Частица *ага* употребляется в позиции вопрос-ответ-частица чаще остальных и в более типичных, соответствующих привычной рутине контекстах, причем как в качестве самостоятельной реплики, возвращающей очередность говорить собеседнику (1), так и в составе более длинной реплики (2).

(1) З — клиент, молодая женщина, О — регистратор, пожилая женщина

1. О: семисят три, (.) какие жалобы скажите.
2. (.)
3. З: затруднённое очень дыхание, [не] хватает воздуха ему.
4. О: [**ага**,]
5. О: затруднён[ное] (.) дыхание, (.) а в- скорую не вызывали?
6. З: [((сглатывает))]
7. З: нет, не вызывали. (.) вот хотели врача вызвать.
8. О: ну вы (.) смотрите по состоянию если мало ли ухудшение будет [то ско]рую вызовите ладно?
9. З: [угу.]

В примере представлен отрывок фазы сбора данных для регистрации вызова. О записывает возраст пациента (начало строки 1) и спрашивает о причине вызова (это необходимо как для ориентировки врачу, так и для отсеивания недостаточно важных показаний, см. (9)). Первая фраза ответа интонационно оформлена как незаконченная; сразу же после нее (в наложении на продолжение реплики) О произносит *ага*, маркируя полученную информацию как релевантную. Слегка нисходящая, проецирующая продолжение интонация *ага* эксплицирует согласие О с намерением З продолжать. Сразу после завершения ответа (маркированного окончательным понижением интонации) О повторяет его первую часть (с. 5), отделяя слова паузами (очевидно, именно эту, маркированную частицей *ага* часть ответа, она записывает), а затем переходит к выяснению серьезности положения.

Частица *а*, также встречающаяся в позиции вопрос-ответ-частица как одиночная реплика, в отличие от нейтрального *ага* почти всегда вопросительно интонирована (единственное в корпусе исключение — пример 5) и вызывает исправление предыдущей реплики, т. е. ее повтор или уточнение.

(2) З — клиент, женщина средних лет, О — регистратор, молодая женщина

1. З: наш доктор э (.) семейный Васнецов,
2. О: угу
3. (.)
4. З: вот, и нам бы надо на дом вызвать.
5. О: а што случилось [такое?]
6. З: [у меня] тут сын пришёл со школы.
7. (0.5)
8. О: **а**?
9. (.)
10. З: сын со школы пришёл больн[ой, те]мпе[ратура.]

11. О: [**аха**.] [температу]ра да? (.) сколько температура у него?=
12. З: =тридцать девять.
13. О: .х адрес какой?

Отрывок начинается с вводной секвенции к просьбе зарегистрировать домашний визит. О реагирует на вводное высказывание нейтрально-плоско интонированной частицей *угу* (с. 2), что З воспринимает как возвращение очередности говорить и переходит к просьбе, отмечая завершение вводной секвенции частицей *вот*. На строке О задает вопрос о причине вызова (с. 5), окончание которого пересекается с началом ответа на с. 6. Ответ интонационно завершен, однако для О недостаточен, т. к. О ждет полсекунды («пауза замешательства»), а затем возвращает очередность говорить З при помощи краткой вопросительно интонированой частицы *а*. З воспринимает реплику *а?* как побуждение повторить и уточнить ответ, т. к. повторяет свое изначальное высказывание, восполняя потенциальную недослышанность из-за перекрывания реплик на с. 5–6, и добавляет значимое дополнение. После произнесения опознаваемой части слова *больной* О отмечает достаточность и релевантность исправления при помощи частицы *ага* со нисходящей, завершенной интонацией. Одновременно с *ага* З дополняет свое высказывание словом *температура*. Через краткий промежуток в микропаузу (один слог) после частицы *ага* О переспрашивает недосказанное слово и, услышав окончание слова, задает следующий вопрос, на который получает быстрый и краткий ответ.

Можно отметить, что *ага* произносится в точке, в которой предшествующая информация отмечается как достаточная (см. тж. Герасименко 2008) — чего нельзя сказать об *а* и *гм*. Прототипное использование частицы *а* — маркер неожиданности. Однако контексты употребления *а* в вопросо-ответных секвенциях — реакция не на всякую неожиданную информацию, но прежде всего на информацию, которая для полного понимания нуждается в исправлении или уточнении. Частица *а* в начале сложных реплик обычно соседствует с вопросом на понимание или прояснение предыдущей реплики, как в (3).

(3) З — регистратор, женщина средних лет, О — больничный электрик, мужчина средних лет

1. З: извините пожалуста это с поликлиники беспокоят.
2. (.)
3. О: да, я слуша[ю.]
4. З: [.х] у нас здесь в-в (.) пришли в-в (.) проводят кажется для компьютера какие-то в провода, (.) и вот им нужен электрик.
5. (0.4)
6. О: **а**, на поликлинику элект[рик?]
7. З: [да,] на поликлинику.
8. (.)
9. О: так вы: не сюда звоните, вас обслуживает Василь, в инфекционном отделении.

Регистратор поликлиники звонит в больницу с просьбой прислать техника. После вводной идентификационной секвенции (с. 1–3) З озвучивает причину звонка. О реагирует на нее после небольшой паузы замешательства частицей *а*, за которой следует вопрос на прояснение (интерпретационная гипотеза, соединяющая информацию из двух реплик З). В следующей реплике З (с. 7) гипотеза О получает подтверждение, затем О сообщает З, что она ошиблась адресатом.

Частицы *а* и *ага* встречаются также перед сигналом о завершении диалога, однако использование *ага* в них более частотно и предшествует более рутинно-лаконичной завершительной конструкции (4), чем использование *а* (5).

(4) З — клиент, женщина средних лет, О — регистратор, женщина средних лет
1. З: скажите пожалуста как Мартыненко сёдня принимает?
2. О: > сегодня минуточку, < (.) .х с пятнадцати до девятнадцати.
3. З: **а:га**, спасибо.
4. О: пожалуйста.

В ответ на вопрос о времени приема З получает обычный рутинный ответ и принимает его к сведению при помощи слегка растянутого *ага* и формулы благодарности; реплика О *пожалуйста* закрывает рутинную завершающую секвенцию и завершает диалог.

(5) З — медик, женщина средних лет, О — регистратор, женщина средних лет
1. З: он после обеда буит?
2. О: .хх ну наверно, он сечас в дневном стационаре.
3. (.)
4. З: **а**, (.) ну всё [{-}, понятно, спасибо.]
5. О: [угу, пожалуйста.]

До начала отрывка З узнала, что главврач уехал; в первой реплике она задает вопрос о времени его возвращения и получает неточный, хотя и максимально подробный ответ. После микропаузы З произносит частицу *а* со слегка нисходящей интонацией и после еще одной микропаузы сообщает О о своей готовности завершить разговор в форме отказа от дальнейших расспросов. О принимает и подтверждает этот сигнал о завершении при помощи частицы *угу*; разговор завершается формулами *спасибо* и *пожалуйста* в наложении друг на друга.

Единственный случай, в котором одиночное *а* имеет нисходящую интонацию и не сопровождается дальнейшим коммуникативным вкладом говорящего, представляет собой отклонение от обычной коммуникативной модели, т. е. вызывает конфликт ожиданий говорящих, проявляющийся в дальнейшем ходе разговора:

(6) З — клиент, молодая женщина, О — регистратор, женщина средних лет
1. З: =а как Мартыненко принимает?

2. (.)
3. О: та:к, доктор Мартыненко, .xx она сегодня будет с трёх до семи.
4. (.)
5. З: **а:**,
6. (1.0)
7. О: мм пожалуста.
8. З: а (.) мне врача бы вызвать, у меня у ребёнка температура.
9. (0.4)
10. О: ну так вы ТАК СРАЗУ И ГОВОРИТЕ што вам нужен?
11. (0.9)
12. З: {-} у неё трицать восемь и девять. (0.3) три годика ребенку. ((жалобным тоном))

Клиент задает вопрос о времени приема врача, выслушивает ответ и после микропаузы произносит немного растянутую частицу *а* со слегка нисходящей интонацией. За ней следует долгая пауза длиной в секунду, свидетельствующая о сбое в смене говорящих: очевидно, О истолковывает предыдущую частицу *а* как готовность З взять коммуникационную инициативу на себя; это подтверждает и маркер хезитации *мм* перед попыткой О инициировать завершение разговора (с. 7). З не поддерживает завершение, высказывая робкое пожелание вызвать врача и приводя причину вызова. После паузы замешательства (с. 9) О эксплицирует нарушение коммуникативной модели со стороны З, повышенным тоном указывая, что З следовало бы высказать свое пожелание *сразу*, на более раннем этапе. Это высказывание О (с. 10) остается грамматически и интонационно незавершенным, что в свою очередь вызывает длительную паузу в смене говорящих. З прерывает паузу на последней строке, начиная перечислять показания к вызову врача, тоном подчеркивая их серьезность (и, возможно, извиняя свою коммуникативную неуспешность).

Частица *гм* встречается в корпусе исключительно в нетипичных, выпадающих из общей рутины контекстах: там, где предшествующая реплика или последовательность реплик оценивается собеседником как не отвечающая его (нормативным) ожиданиям: *гм* предваряет негативную оценку незнания клиентом важной информации (7) или указание клиенту на недостаточность симптомов для визита врача. Интересно, что частица *гм*, как и сами негативные оценочные реплики, используется лишь регистраторами; клиенты, также могущие быть недовольными отказом прислать врача или недостаточностью полученной информации, оценок в разговоре не высказывают и обычно выражают свое недовольство раздраженным тоном уточняющих вопросов.

(7) З — клиент, женщина средних лет, О — регистратор, женщина средних лет.
1. О: так, телефон домашний у них есть?
2. (0.5)

3. З: не (.) дали мне ничего.
4. (.)
5. О: гм, очень плохо. (.) .хх ну всё тогда, ждите.

В отрывке приведено завершение диалога, в котором врача ребенку вызывает знакомая его матери, не будучи в курсе симптомов и того, как попасть в подъезд (не входит в отрывок). На с. 1 О задает очередной вопрос, ответа на который З не знает (и формулирует ответ, подчеркивая ответственность тех, кто не снабдил ее информацией), на что О после микропаузы реагирует слегка нисходящей частицей *гм* и оценкой *очень плохо*. Затем О завершает регистрацию, подтверждая действительность вызова (*ждите*).

Выводы

Анализ демонстрирует, что для определения коммуникативных различий частиц важно учитывать временную соотнесенность их со стимулом — для *ага* характерны быстрота реакции и отсутствие заминки, в то время как *а* и *гм* обычно отделены от стимула «паузой замешательства».

Кроме того, важен не только предшествующий, стимульный, но и последующий, интерпретационный контекст. Функциональные различия частиц проявляются не только в интерпретации собеседников, но и в продолжении реплик самими говорящими: *гм* в начале реплики предваряет оценку стимульной реплики, *а* и *ага* предшествуют уточняющим вопросам. Вопросы, следующие за *а,* направлены на ликвидацию неясностей в реплике-стимуле, а вопросы после *ага* направлены на дальнейшее детализирование или развитие темы (разговора). Частицы *а* и *ага* могут также предшествовать сигналу говорящего о завершении разговора. В случае *ага* это завершение более лаконично и рутинно (*ага, спасибо*); лексема *а* сочетается с отказом от дальнейших распросов (*а, ну все тогда; а, ну ладно*), который косвенно указывает на сохраняющуюся (частичную) непроясненность.

В позиции одиночной реплики, возвращающей очередность говорить собеседнику, *ага* в корпусе нейтрально интонировано и побуждает собеседника продолжать, в то время как *а* маркировано вопросительной интонацией и побуждает собеседника повторить или дополнить уже сказанное.

Мы надеемся, что результаты анализа дополнят существующие описания частиц с точки зрения их осознанного использования в русском диалоге. Исследования различий использования частиц в других спонтанных контекстах (в бытовой речи, в спонтанном построении нарративов, в институциональном общении другого рода) позволят еще лучше и точнее выявить их инвариантные коммуникативные характеристики.

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:
вдох	.хх
убыстрение темпа	> текст <
сомнительный отрезок	{текст}
неразборчивый отрезок	{--}
повышение громкости	ТЕКСТ
комментарии	((текст))

Литература

1. *Большой* толковый словарь русского языка. Глав. ред. Кузнецов С. А., Санкт-Петербург: Норринт, 2006.
2. *Герасименко О. А.* 2008. Два значения, две языковых единицы? Ага в спонтанном диалоге — Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям Диалог 2008, Москва, 103–108.
3. *Гришина Е. А.* 2009. К вопросу о соотношении слова и жеста (вокальный жест О в устной речи) — Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям Диалог 2009, Москва, 80–90.
4. *Гришина Е. А.* 2010. Вокальный жест А в устной речи — Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям Диалог 2010. Москва: Наука, 102–112.
5. *Добрушина Н. П.* 1998. Семантика междометий в реактивных репликах -Вестник Московского университета 2: 136–145.
6. *Златев И.* 2004. Вопрос о разграничении междометий и частиц в болгарском и в русском языкознании — Ежегодные лингвистические чтения. 15 ноября 2003 года. В. Тырново, 2004, http://www.russian.slavica.org/printout1195.html (дата обращения: 05.02.2012).
7. *Иомдин Б. Л.* 2004. Междометия догадки в русском языке — Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям Диалог 2004. Москва, http://www.dialog-21.ru/Archive/2004/Iomdin%20B.htm (дата обращения: 09.02.2012).

8. *Исупова О.* 2002. Конверсационный анализ: представление метода — Социология 4М, № 15, http://www.isras.ru/files/File/4M/15/Isupova.pdf (дата обращения: 05.02.2012).
9. *Русская* грамматика 1980: Русская грамматика. Т.I- II, Москва, 1980–1982.
10. *Словарь* современного русского литературного языка: В 17-ти т. под ред. В. И. Чернышёва — Москва, Ленинград: Изд-во АН СССР, 1948–1965.
11. *Шаронов И. А.* 2008. Междометия в речи, тексте и словаре. Москва: РГГУ.

АВТОДЕЙКСИС: ОСНОВНЫЕ ТИПЫ И ЗНАЧЕНИЯ[1]

Гришина Е. А. (rudi2007@yandex.ru)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

В статье анализируются типы автодейксиса, т.е. типы жестов, которые используются говорящим при указании на самого себя. Исследуются также основные языковые факторы, влияющие на выбор типа жестикуляционного автоуказания.[2]

Ключевые слова: дейксис; автодейксис; указательные жесты; соотношение слова и жеста; Мультимедийный русский корпус (МУРКО)

1. Введение: парадоксы автодейксиса

В ходе живого диалогического общения, как и в ходе устного монолога указание на самого себя является достаточно частым жестом — не менее частым, чем указание на любые внешние объекты, лежащие вне актуальной в данный момент коммуникативной сферы. Судя по частотности этого жеста, он представляет собой не исключение, а правило, т. е. является стандартным дейктическим сопровождением устной речи.

При этом, однако, смысл и назначение этого жеста совершенно не очевидны и, если вдуматься, не вполне понятны. Неясно, почему местоимения *я* и *мы*, не являясь демонстративами, регулярно сопровождаются указательными жестами.

Далее, оказывается совершенно не очевидным ответ на вопрос — на что именно указывает человек, когда он осуществляет жестикуляционный автодейксис, — на себя как личность? на себя как тело? или существуют еще какие-то возможности? Где находится тот объект, на который производится жестикуляционное автоуказание, и какова природа этого объекта?

И наконец, наблюдения показывают, что указующая рука в случае автодейксиса может иметь разную конфигурацию. Это тоже является своего рода парадоксом — ведь в случае автодейксиса и объект указания, и субъект указания являются одной и той же постоянной сущностью. Так какие же факторы

[1] Исследование проведено при поддержке программы Президиума РАН «Корпусная лингвистика», а также грантов РФФИ № 10-06-00151 и № 11-06-00030.

[2] Доклад публикуется в сокращенном варианте. Полная версия доступна на сайте конференции «Диалог»

влияют на выбор конфигурации указующей руки? или выбор конфигурации совершенно свободен и не несет какого-то дополнительного смысла?

Итак, с одной стороны, мы имеем достаточно частотный дейктический жест, настолько обычный, что ни сам говорящий, ни слушающий, ни сторонний наблюдатель чаще всего не обращают на него внимания. С другой стороны, мы имеем местоимение первого лица, которое имеет кристально ясную семантику (*я* — тот участник коммуникативного акта, реплика которого звучит в данный момент). И при этом совершенно непонятен смысл соединения, сосуществования этих двух абсолютно стандартных явлений.

В данной статье мы хотели бы, во-первых, перечислить основные жестикуляционные типы автодейксиса, а во-вторых, предложить некоторые версии, объясняющие перечисленные выше парадоксы.

2. Где находится *я*?

Предметом рассмотрения в нашей работе будут чуть более 300 примеров автодейксиса, отобранных из Мультимедийного русского корпуса[3]. Разумеется, объем материала не слишком велик, но, с нашей точки зрения, достаточен для предварительных выводов.

Будем считать, что говорящий знает, где именно находится *я*, и при автодейксисе указывает именно на это место. Тогда точкой *я*, в соответствии со статистическими данными, следует считать небольшую зону в районе диафрагмы или чуть выше (на рис. 1 именно в точку *я* ранит святого Себастьяна одна из стрел). Именно в точку *я* направлена основная часть жестикуляционных автодейксисов.

Рис. 1. Точка *я*

[3] О МУРКО см. [Гришина 2009] и инструкцию на сайте www.ruscorpora.ru/search-murco.html.

3. Основные понятия и термины

3.1. Встроенные и независимые жесты

В соответствии с классификацией, данной в работе [Magno-Caldognetto, Poggi 1997: 34], выделяются следующие основные режимы (в работе — *модусы*) соотношения жеста и слова:

1) повтор, когда жест дублирует значение, которое языковыми средствами выражено в высказывании:

(1)

Речевой ряд	*А какая разница, кто вам скажет — он и ли*	*я?*
Жестовый ряд		автодейксис, указание ОЛ

А. Коренев и др. Вас вызывает Таймыр, 1970

(2)

Речевой ряд	*Это*	*я разбил!*
Жестовый ряд		автодейксис, указание УП

А. Сурикова, В. Еремин. Хочу в тюрьму, 1998

2) дополнение, когда жестикуляция добавляет в сопровождающее высказывание некоторое значение, при том что высказывание само по себе является вполне завершенным:

(3)

Речевой ряд	[Лебедев] *Ибо нищ и наг, и атом в коловращении людей. И кто почтит*	*Лебедева?*
Жестовый ряд		автодейксис, указание УП
Значение		‘Лебедев — это я’

В. Бортко. Идиот, 2003

(4)

Речевой ряд	*У нас в клубе принято*	*джентльменам*	*верить на слово.*
Жестовый ряд		автодейксис, указание ОЛ	
Значение		‘джентльмены — это мы’	

3) замещение, когда жест используется вместо слова или словосочетания, а высказывание без жеста является незавершенным или неполным:

(5)

Речевой ряд	*В этом прежде всего заслуга*	[пауза]	*его родных.*
Жестовый ряд		автодейксис, указание ОЛ	
Значения		‘моя’	

М. Захаров, Г. Горин. Тот самый Мюнхгаузен, 1979

(6)

Речевой ряд	*Высек*	[пауза]	*и тем самым запечатлел.*
Жестовый ряд		указывает на Рогожина (УП), затем указывает на себя (автодейксис УП)	
Значение		‘он меня’	

В. Бортко. Идиот, 2003

4) конфликт, когда значение жеста противоречит значению высказывания (встречается достаточно редко, к дейктическим жестам отношения не имеет, далее учитываться не будет):

(7)

Речевой ряд	*— Сеня, ты уже дошел до кондиции?* *— До какой? — До нужной?*	*— Нет.*
Жестовый ряд		утвердительно кивает головой

Л. Гайдай и др. Бриллиантовая рука, 1968

5) автономность, когда жест выражает значение, полностью отсутствующее в параллельном речевом ряде:

(8)

Речевой ряд	*Не-не,*	*это*	*полный бред!*
Жестовый ряд		автодейксис, указание ОЛ	

Д. Дьяченко и др. День радио, 2008

6) И наконец, последним способом, которым жестикуляция может сочетаться с параллельным речевым рядом, является режим **пантомимы**, когда речь отсутствует, а жест выступает как самостоятельное высказывание.

Далее мы будем называть ***встроенными автодейксисами*** те автоуказания, которые используются в режиме повтора, дополнения и замещения (пп. 1–3 выше), т. е. встроены в высказывание, которое сопровождают, а ***независимыми автодейксисами*** — автоуказания, которые используются в автономном режиме (п. 5).

Среди встроенных жестов полезно выделить ***я-синхронные жесты***, ударная стадия которых совпадает с местоимением *я* или его производными. Все остальные типы соотношения местоимения *я* и автодейксиса (независимые и встроенные, но не *я*-синхронные) мы будем называть ***асинхронными***.

3.2. Базовые автодейксисы

В работах [Гришина 2012а, б] нами на обширном материале были рассмотрены русские дейксисы *вот, вон, здесь, тут, там, тот, это, этот*

и проанализированы те ручные дейктические жесты, которые их сопровождают. В результате было показано, что разные аспекты значения исследованных дейксисов отражаются в разных конфигурациях жестикулирующей руки. Выяснилось, что значение дейксиса и тип речевого акта, содержащего в себе дейксис, влияют, с одной стороны, на выбор органа указания — указательный палец (УП) или открытая ладонь (ОЛ). С другой стороны, обнаружилась связь между значением дейксиса/типом речевого акта — и ориентацией ладони (вверх или вниз; вертикальная ориентация ладони оказалась референциально и прагматически немаркированной, т. е. не имеющей собственного значения).

Рис. 2. Автоуказание пальцем

Рис. 3. Автоуказание открытой ладонью

Поскольку при автодейксисе ориентация ладони практически всегда имеет вынужденный характер, то для анализа автоуказания, по-видимому, следует различать только указание пальцем (рис. 2) и указание открытой ладонью (рис. 3) — далее УП и ОЛ, — и не учитывать ориентацию ладони. Эти две конфигурации указующей руки мы будем назвать ***базовыми автодейксисами***.

3.3. Зона действия жеста

Нам понадобится такое понятие, как зона действия жеста. Как известно, обычно жест осуществляется в три стадии — подготовительная стадия (экспозиция), ударная часть (stroke; основная смысловая часть жеста) и ретракция. Из этих трех частей только ударная является обязательной. Кроме того, очень часто в жест добавляются предударное и послеударное удержания (holds) — задержки в осуществлении жеста, когда активный орган фиксируется в стадии экспозиции (предударное удержание) или в стадии удара (послеударное удержание). ***Зоной действия жеста*** мы будем считать ту часть сопутствующей жесту фразы, которая соответствует ударной части жеста и послеударному удержанию, если оно наличествует. Так, в примере (8) зона действия жеста совпадает со словом *это*, в примере (1) — со словосочетанием *я разбил*.

4. Базовые автодейксисы

Базовые автодейксисы ОЛ и УП заметно различаются с точки зрения их отношений с сопровождающей речью. Табл. 1 показывает распределение встроенного и независимого режима употребления этих автодейксисов.

Таблица 1

Тип употребления / Тип автодейксиса	Встроенный автодейксис	Независимый автодейксис
ОЛ	149 (75%)	50 (25%)
УП	62 (98%)	1 (2%)
χ^2=16,9[2], распределения достоверны, параметры связаны		

Как видим, четверть всех автодейксисов ОЛ употребляется в независимом режиме, т. е. в сопровождающем жестикуляцию тексте нет ни самого слова *я*, ни его производных, ни форм, в которые может быть включена сема 1-го лица (глаголы, слова типа *сам, собственный*), ни личных имен говорящих. Один пример такого рода употреблений уже был приведен выше (8), вот еще несколько:

(9)

Речевой ряд	*Нет, Юрий*	*Федорович!*	*Почему эта тема хуже?*
Жестовый ряд		автодейксис ОЛ	

Д. Дьяченко и др. День радио, 2008

(10)

Речевой ряд	*С твоими*	*поучениями ты бы*	*выступал у себя на рынке!*
Жестовый ряд		автодейксис ОЛ	

Э. Рязанов, Э. Брагинский. Берегись автомобиля!, 1966

(11)

Речевой ряд	*…если бы не увидел, что она оказала тебе*	*высший почет*	*проводив тебя до самых*	*дверей класса*
Жестовый ряд		автодейксис ОЛ		автодейксис ОЛ

Г. Казанский, Л. Лагин. Старик Хоттабыч, 1956

Единственный же пример, в котором в независимом режиме употребляется автодейксис УП, не вполне показателен (12):

[4] При каждой таблице с числовыми данными мы даем значение соответствующего данному распределению χ-квадрата (χ^2). Это позволяет оценить степень достоверности приведенных в таблице данных.

(12)

Речевой ряд	*Все*	*люди*	*любят редких животных,*	*а туристов в Конго, я уверен, не все!*
Жестовый ряд	назидательный жест *ткнуть пальцем*	автодейксис УП	назидательный жест *ткнуть пальцем*	

Д. Дьяченко и др. День радио, 2008

В примере (12) мы, вероятнее всего, имеем дело с довольно частотным явлением, *аккомодацией жестов*, когда в жестикуляционном континууме ретракция предшествующего жеста влияет на ударную часть последующего. Как видим, в примере (12) автодейксис находится между двумя жестами *ткнуть пальцем* в значении назидания, так что здесь конфигурация руки в автодейксисе (УП) может иметь зависимый, индуктивный характер. А если это так, то мы можем утверждать, что на нашем материале ни разу не встретился пример автономного автодейксиса УП, при том что автономный автодейксис ОЛ чрезвычайно частотен.

Иными словами, для того, чтобы говорящий указал на себя указательным пальцем, в сопровождающем высказывании, непосредственно относящемся к этому автодейксису, обязательно должна присутствовать словесная отсылка к говорящему в любой форме (местоимение, глагол, имя говорящего). Напротив, автодейксис ОЛ свободен от таких ограничений и может быть употреблен и в тех случаях, когда в сопровождающем высказывании нет ни одного маркера первого лица.

Здесь естественным образом возникают два вопроса: 1) почему говорящий использует жестикуляционное указание на себя, если в звучащем высказывании нет показателей ‘я’?, 2) почему в том случае, если в сопровождающем высказывании нет показателей ‘я’, говорящий при автодейксисе выбирает указание ОЛ, а не УП?

Начнем с первого вопроса. Тот факт, что автодейксис УП жестко связан с наличием местоимения *я* и сходных лингвистических явлений в текстуальной части высказывания, позволяет нам считать, что автодейксис УП относится к ‘я’, входящему в диктум, т. е. в содержательную зону значения данного высказывания. Иными словами, автодейксис УП указывает на ‘я’, которое прежде всего является топиком данного высказывания, и лишь во вторую очередь маркирует *сам акт высказывания*.

Если это так, то мы можем предположить, что автодейксис ОЛ обозначает **не** диктальное ‘я’. И действительно, каждое высказывание может включать или не включать в себя диктальное местоимение первого лица (*я* или *мы*), т. е. показатель первого лица как предмет речи, но *обязательно* включает в себя перформативное первое лицо (‘я’), поскольку каждое высказывание имеет при себе перформативный префикс, обозначающий тот тип иллокуции, который в данный момент осуществляет говорящий (<Я говорю>, <Я спрашиваю>, <Я обещаю> и т. д.). В большинстве высказываний перформативный префикс функционирует в скрытом, подразумеваемом виде, но иногда формулируется и на поверхностном уровне:

(13) ***А я вам говорю**, Герснер, учтите…*
*Но **у меня еще один особо важный вопрос**…*
***Это я спрашиваю**, что случилось.*
***Это я вас хочу спросить**, почему…*
***И я говорю**, понимаете, вот…*
***Я прошу** понять меня правильно…*

Таким образом, сема ‘я’ присутствует в любом высказывании, и если ее нет на уровне содержания, диктума, то она обязательно имеется в составе скрытого перформатива.

Итак, мы различаем диктальное и перформативное ‘я’ и считаем, что автодейксис УП апеллирует к диктальному ‘я’, в то время как автодейксис ОЛ апеллирует к ‘я’ перформативному. Режим встроенного автодейксиса в значительной степени нейтрализует эти различия, но они выходят на первый план в независимом режиме жестикуляции, при котором используется исключительно автодейксис ОЛ; автодейксис УП употребляется только во встроенном режиме.

Если автодейксис УП относится к диктальному ‘я’, естественно ожидать, что он должен реагировать на различия в типах речевых конструкций, включающих в себя ‘я’. Будем различать контрастные и нейтральные контексты. В контрастных контекстах местоимение *я* попадает под эмфатическое выделение (примеры (1), (2)) или сочетается с выделяющими частицами *-то, вот*. Нейтральными мы будем считать все остальные контексты. Табл. 2 демонстрирует соответствующие данные.

Таблица 2

Тип контекста **Тип автодейксиса**	**Контрастный**	**Нейтральный**
ОЛ	55 (42%)	76 (58%)
УП	25 (66%)	13 (34%)
χ^2=6,7, распределения достоверны, параметры связаны		

Как видим, автодейксис ОЛ слабо тяготеет к нейтральным конструкциям. Напротив, автодейксис УП отчетливо предпочитает контрастные контексты (две трети случаев использования автодейксиса УП). Таким образом, речевая конструкция (противопоставление контрастных и нейтральных контекстов) имеет значительное влияние на диктальный автодейксис УП, а перформативный автодейксис ОЛ затрагивает в существенно меньшей степени.

Интересно проследить, связаны ли между собой зона действия жеста (см. раздел 3.3) и тип автодейксиса (ОЛ или УП). Будем считать, что автодейксис связан со **словом**, если в зону действия жеста попадает только одна словоформа (иногда — вместе с примыкающими к ней клитиками), и со **словосочетанием**, если в зону действия жеста попадает больше одного полнозначного слова. При этом выведем из рассмотрения однословные реплики (типа *Мне-то?,*

Представляете?), в которых диапазон зоны действия жеста не имеет вариантов. Данные представлены в табл. 3.

Таблица 3

Зона действия жеста **Тип автодейксиса**	**Слово**	**Словосочетание**
ОЛ	84 (45%)	102 (55%)
УП	40 (66%)	21 (34%)
χ^2=7,7, распределения достоверны, параметры связаны		

Как видим, в случае автодейксиса ОЛ выбор между зоной действия жеста длиной в одну словоформу и длиной в словосочетание практически равновероятен (с небольшим предпочтением словосочетания). Напротив, в двух третях случаев употребления автодейксиса УП последний задает зону действия длиной в одно слово.

Объясняется это, с нашей точки зрения, тем, что длительность автодейксиса УП, ориентированного на ‘я’ как на часть диктума, определяется тем, что ‘я’ в составе диктума обычно представлено одной словоформой (будь то сами местоимения *я, мы, мой, наш*, или личные имена говорящих, глагольные словоформы первого лица). Напротив, автодейксис ОЛ, ориентированный на скрытый перформатив высказывания, не связан с какой-либо конкретной словоформой, а длится столько, сколько длится, с точки зрения говорящего, соответствующий перформатив. Таким образом, автодейксис УП можно считать сегментным, а автодейксис ОЛ — суперсегментным автодейксисом. Если обозначить границы зоны действия жеста фигурными скобками, то мы заметим, что зона действия жеста, подобно интонационному контуру, в значительном числе случаев хорошо совпадает с синтаксическими границами фразы:

(14) *Только я {если ты не возражаешь} я бы все-таки взяла вот эти…*
Я тебя разбудил {извини пожалуйста} я не подумал
{Почему я должна ехать} на метро?
Эта твоя независимость {ну зачем она мне?}
А ты знаешь {во сколько я ложусь?}

Если принять предложенное разделение автодейксисов на диктальный и перформативный, можно задаться вопросом, почему для перформативного дейксиса выбрано указание ОЛ, а для диктального — указание УП.

Наш анализ системы ручных указательных жестов [Гришина 2012а, б] показал, что, во-первых, для указания УП характерна трактовка объекта указания как **точечного** объекта стабильной структуры, в то время как в случае указания ОЛ объектом указания является ситуация, **событие**, расположенное в трех измерениях и развивающееся во времени. Во-вторых, указание УП обычно используется в идентификациях типа *X это Y*, в которых обе части

идентификации (*X* и *Y*) являются понятийными и, следовательно, речевыми объектами. Напротив, указание ОЛ тяготеет к идентификациям структуры *Вот X*, в которых *X* является понятием и, соответственно, **словом**, а введенный с помощью данной идентификации объект является конкретным предметным или событийным (т. е. **внеречевым**) примером реализации данного понятия (*это*-идентификация осуществляет категоризацию объекта указания, а *вот*-идентификация — экземплификацию категории).

Отметим, что диктальное ‘я’ как составная часть высказывания может рассматриваться как **точечный речевой** объект стабильной структуры. Напротив, перформатив ‘я говорю’ логично трактовать как **внеречевое событие**, развивающееся во времени. Если принять такую точку зрения, то обе упомянутые выше закономерности в выборе конфигурации указующей руки в приложении к автодейксису должны приводить к тому, что автодейксис УП будет использоваться для диктального автоуказания, а автодейксис ОЛ — для перформативного.

Именно этим, с нашей точки зрения, вызван тот факт, что перформативный автодейксис ОЛ не употребляется в режиме пантомимы: на всем нашем материале (204 случая автодейксиса ОЛ) последний ни разу не был употреблен как самостоятельное высказывание, без сопровождающей речи. Что касается автодейксиса УП, то здесь ситуация иная — из 47 случаев его употребления в двух случаях автоуказание пальцем использовалось в режиме пантомимы, и в обоих случаях «переводилось» с жестикуляционного языка на словесный совершенно однозначно: ‘это **я** Р’, где Р, предикат, понятен обоим собеседникам из предшествующей ситуации, а жестикулирующий субъект своей пантомимой утверждает, что это именно он является субъектом, причиной, виновником или каким-то другим актором данного предиката. Очевидно, что раз автодейксис ОЛ апеллирует к перформативу, то он в обязательном порядке должен быть связан с событием высказывания, поскольку перформатива без акта высказывания не существует. Напротив, автодейксис УП, связанный не с событием речи, а с ее содержанием, может передавать это содержание и без речи, в режиме пантомимы.

В заключение раздела попытаемся ответить еще на один вопрос: зачем вообще говорящему указывать на перформативное ‘я’? Логика указания на диктальное ‘я’ достаточно прозрачна: в режиме замещения и дополнения (см. выше, разд. 3.1) такое указание проясняет валентный и содержательный состав сопутствующего высказывания; в режиме повтора автодейксис либо дублирует содержание высказывания[5], либо добавляет в него те или иные обертоны (в частности, автодексис УП, апеллирующий к диктальному ‘я’, может добавлять к нему контрастное выделение, подчеркивание, отсутствующее в других модусах данного высказывания).

Но в чем смысл перформативного автодейксиса?

В работе [Гришина 2012б] мы уже писали о том, что для русской дейктической жестикуляции чрезвычайно характерна жестикуляционная фраза, т. е.

[5] Дублирование содержательной части устного высказывания разными модальными средствами, — фонетическим, интонационным, жестикуляционным, — явление достаточно частотное.

последовательность жестов, структуры ‘ты–он–ты’, которая может быть переведена примерно следующим образом: ‘обрати внимание’ — ‘фиксация объекта’ — ‘обрати внимание’ (одно из двух ‘ты’ часто опускается), ср. примеры (15)–(17).

(15)

Речевой ряд	*Вот!*	*Признали свою*	*вину!*
Жестовый ряд	направляет указательный палец на третье лицо	переводит указательный палец на собеседника	снова переводит указательный палец на третье лицо
Значение	фиксация объекта	‘обрати внимание’	‘на этот объект’

А. Рогожкин. Операция «С Новым Годом!», 1996

(16)

Речевой ряд	[*Вытирает мороженое с лица, смеется*]	*Хороший мальчик…*
Жестовый ряд	направляет указательный палец на собеседника	переводит указательный палец на третье лицо
Значение	‘обрати внимание’	‘на этот объект’

Л. Гайдай и др. Бриллиантовая рука, 1968

(17)

Речевой ряд	*Да он это!*	*Он! Он самый!*
Жестовый ряд	направляет указательный палец на собеседника	переводит указательный палец на третье лицо
Значение	‘обрати внимание’	‘на этот объект’

Л. Кулиджанов, Н. Фигуровский. Когда деревья были большими, 1961

Мы предполагаем, что перформативный автодейксис используется по сходной схеме: говорящий обращает внимание слушающего на тот факт, что «**я (говорящий) говорю**». Таким образом, можно предположить, что перформатив, лежащий в основе любого устного высказывания, включает в себя не только глагол говорения в первом лице единственного числа (‘говорю’, ‘прошу’, ‘требую’, ‘обещаю’ и др.), но и апелляцию к слушающему по схеме ‘ты–я–ты’, и переводится примерно следующим образом: ‘обрати внимание’ — ‘говорю’ — ‘тебе’.

Если это так, то реплики, которые сопровождаются перформативным автодейксисом ОЛ, должны быть в большей степени ориентированы на адресата, чем реплики, содержащие диктальный автодейксис УП.

Ориентация реплик на адресата может быть выражена разными способами.

1) Реплика может быть провоцирующей[6], т. е. нацеленной на обязательную реакцию адресата. Такими репликами, прежде всего, являются вопросы и императивы самой разной структуры.

[6] О провоцирующих репликах и специальных схемах глазного поведения говорящего и слушающего на их границах по сравнению с нейтральными репликами см. [Гришина 2011а].

2) Реплика может содержать обращение.

3) И наконец, ориентация высказывания на адресата может выражаться не на словесном, а на жестикуляционном уровне, а именно, она может сопровождаться жестикуляционной фразой структуры ‘ты–я–ты’ (одно из двух ‘ты’ может отсутствовать), см. пример (18) и рис. 4, на котором говорящий сначала делает движение открытой ладонью к адресату (первое ‘ты’), затем осуществляет автодейксис ОЛ (‘я’), а затем повторяет движение открытой ладонью к адресату (второе ‘ты’).

Рис. 4. Жестикуляционная фраза *'ты–я–ты'*

(18)

Речевой ряд	*Голодать будут*	*за еду!*
Жестовый ряд	ОЛ к адресату, ‘ты’	автодейксис ОЛ, затем ОЛ к адресату, ‘я–ты’
Значение	‘обрати внимание’	‘говорю–тебе’

О. Фомин и др. День выборов, 2005

Табл. 4 демонстрирует, в какой степени высказывания, ориентированные на адресата, связаны с автодейксисами ОЛ и УП.

Таблица 4

Ориентация высказывания **Тип автодейксиса**	**На адресата**	**Нейтральная**
ОЛ	138 (71%)	56 (29%)
УП	18 (40%)	27 (60%)
χ^2=17,6, распределения достоверны, параметры связаны		

Как видим, при автодейксисах ОЛ и УП нейтральная ориентация высказывания и ориентация на собеседника распределены крестообразно: как мы и ожидали, для автодейксиса ОЛ характерна ориентация на собеседника (71%), для автодейксиса УП — нейтральная (объективная) ориентация (60%).

Если посмотреть на типы ориентации более детально, то мы заметим, что в вопросах и императивах (т. е. в провоцирующих высказываниях) автодейксис ОЛ встречается в 34% случаев, а автодейксис УП — только в 15%, т. е. в два раза реже; автодейксис УП вообще не встречается на фоне обращений, а для

автодейксиса ОЛ такое сочетание зафиксировано (5 примеров, один из них см. выше, пример (11)); жестикуляционная фраза ‘ты–я–ты’ с автодейксисом ОЛ встречается в 51 % случаев и только в 29 % случаев — в сочетании с автодейсисом УП, т. е. разница почти в два раза.

Таким образом, мы считаем, что перформативный автодейксис ОЛ предназначен для того, чтобы привлечь внимание собеседника к событию высказывания, в то время как диктальный автодейксис УП прежде всего ориентирован на дублирование или замещение семы ‘я’ как составной части диктума, содержательной зоны высказывания.

5. Заключение

В заключение отметим, что жестикуляционное противопоставление двух основных органов указания — открытой ладони и указательного пальца — хорошо коррелирует с противопоставлением объектов, входящих и не входящих в личную сферу говорящего. Так, в табл. 5 можно видеть распределение органов указания между основными дейктическими словами.

Таблица 5

Дейксис	ОЛ	УП
я	204	45
вон	27	61
вот	125	51
там	53	106
тот	33	66
тут	55	28
χ^2=171,7, распределения достоверны, параметры связаны		

Мы видим, что указание ОЛ характерно для тройки *я–вот–тут*, а указание УП — для тройки *тот–вон–там*, и эти тройки, тем самым, могут рассматриваться не только как речевые, но и как жестикуляционные антонимы. В первой тройке *тут* обозначает событие, развертывающееся в личной зоне говорящего, *вот* приглашает слушающего к восприятию объекта указания как части личной событийной сферы говорящего, а *я* обозначает сердцевину этой личной сферы — *событие речи*.

Литература

1. *Гришина Е. А.* (2009). Мультимедийный русский корпус (МУРКО): проблемы аннотации. Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб. «Нестор-История»: 175–214
2. *Гришина Е. А.* (2011). Грамматика взгляда: направление взгляда как лингвистический фактор. Филология, 1, 50–71
3. *Гришина Е. А.* (2012). *Здесь* и *тут*: корпусной и жестикуляционный анализ полных синонимов. Русский язык в научном освещении, 23 (в печати)
4. *Гришина Е. А.* (2012). Указания рукой как система (по данным Мультимедийного русского корпуса). Вопросы языкознания, 3 (в печати)
5. *Magno-Caldognetto E. A., Poggi I.* (1997). Micro- and macro-bimodality. ESCA Workshop on Audio-Visual Speech Processing (AVSP'97): 33–36

МЕТОДИКА ОЦЕНКИ КОНВЕНЦИОНАЛЬНОСТИ МЕТАФОРИЧЕСКИХ ВЫРАЖЕНИЙ: ОТ ИНТУИТИВИСТСКИХ КРИТЕРИЕВ К ОПЕРАЦИОНАЛЬНЫМ

Гурин Г. Б. (grigory.gurin@uef.fi),
Беликова А. Е. (alex.belikova@gmail.com)

Университет Восточной Финляндии, Йоэнсуу

В статье предпринимается попытка усовершенствования существующих количественных подходов к метафоре в дискурсе и предлагается новая классификация типов метафорических выражений по степени конвенциональности, сохранности исходной образной составляющей. Новизна предлагаемой классификации состоит в минимизации роли интуитивных и интроспективных критериев, которые широко используются в метафорологических исследованиях. Установление значения параметра метафорической креативности опирается главным образом на последовательное использование лексикографических и корпусных данных и в меньшей степени на интуицию исследователя. Предлагаемая классификация сопровождается алгоритмизированной методикой оценки метафорических выражений с точки зрения конвенциональности / креативности. С помощью этой методики возможна количественная оценка степени метафоричности отдельных текстов и их совокупностей. Источником иллюстраций является русский язык.

Ключевые слова: метафора, конвенциональность, креативность, интроспекция и интуиция, корпус, метод, алгоритм

Введение

Общим местом в метафорологии стало представление о том, что мертвые метафорические выражения (далее — МВ) противопоставляются окказиональным, а пространство между этими полюсами заполнено более или менее конвенционализованными МВ.

Существует несколько традиционных и современных классификаций метафор с этой точки зрения, однако интенсивные исследования пока не вылились в формальную методику оценки МВ. Несмотря на отдельные попытки операционального выделения таксонов, четких методик измерения степени конвенциональности, опирающихся на проверяемые признаки и охватывающие все пространство МВ, похоже, пока не существует. Тем не менее имеются работы, прибегающие к количественному измерению степени «интенсивности» или «креативности» метафор в массивах текстов, например (De Landtsheer 1998), (Баранов 2003).

Области применения классификации МВ

Для чего нужна такая классификация? На этот вопрос можно дать три ответа.

1) Традиционно предполагается, что окказиональные сближения двух семантических планов — признак недавнего возникновения МВ, а привычность когнитивного сближения, автоматизм восприятия, незаметность — маркеры старения метафоры, поэтому измерение интенсивности метафор может стать ключом к анализу динамики в эволюции метафорических систем. Исследователь, располагающий методикой установления степени метафоричности, может оценивать инновации с точки зрения перспектив их закрепления в языке и прогнозировать развитие метафорических моделей и систем, что может стать ценным дополнением к каталогизирующему подходу в описании метафор.

2) Измерение конвенциональности МВ оказывается необходимым в прикладных исследованиях, интерпретирующих метафорику, например, как индикатор активности социально-политических процессов. Работы такого типа предполагают кодирование входящей информации с помощью весовых коэффициентов, что неизбежно вызывает вопрос о правилах приписывания количественных значений анализируемым единицам.

3) Формализованная методика анализа степени конвенциональности МВ, эксплицитно отказывающаяся от интуитивистских решений, оказывается единственно адекватной в случае анализа метафор тех языков, полноценным знанием которых исследователь не обладает, например в исторических, типологических и в переводоведческих исследованиях.

Интуитивистские классификации: метафоры *живые* и *мертвые*

Традицию оценки конвенциональности метафор в интуитивистских терминах заложил Ш. Балли (Bally: 193–195) и с тех пор в большинстве метафорологических работ существование разных типов метафор принималось как данность. Термины из этой сферы (метафора *мертвая, языковая, этимологическая, историческая, стандартная, конвенциональная, узуальная, свежая, инновационная, авторская, окказиональная; dead, frozen, entrenched, naturalized, established, novel, innovative, nonce* и т. д.) использовались как сами собой разумеющиеся и не нуждающиеся в операциональном толковании. Потребовалось немало исследовательских усилий для осознания ограниченности этого подхода и для формулирования некоторых эмпирических критериев разграничения разных типов метафор. Однако и в современных работах интуитивистские определения используются неожиданно широко: например, в обобщающей работе (Cruse 2004), в целом часто использующей возможности формального разграничения разных классов единиц, разница между «натурализованными» (naturalized) и «укоренившимися» (established)

метафорами подается как самоочевидная интроспективная данность: первые «настолько утвердились и так глубоко вошли в язык, что говорящие *не чувствуют* [no longer feel], что они представляют собой фигуры речи»; вторые же «входят в ментальный лексикон в виде отдельных единиц, но тем не менее *чувствуется* [felt to be], что это фигуры речи» (Cruse 2004: 197) (курсив наш — *А. Б., Г. Г.*).

Имеются случаи, когда исследователь, нуждающийся в измерении степени конвенциональности МВ, использует в качестве инструмента определения степени узуальности/окказиональности не систему верифицируемых признаков, а компетенцию носителя. Таков подход К. Де Ландтсхеер, реализованный, в частности, в статьях (De Landtsheer 1998) и (Vertessen, De Landtsheer 2005), анализ парламентского дискурса опирается на вычисление коэффициента метафоричности, переменная интенсивности в котором ранжируется с помощью весовых коэффициентов, причем присвоение конкретного значения МВ производится носителями языка на основе их субъективных оценок «странности или необычности метафорического выражения в контексте» и «силы отсылки к буквальному значению».

Принципы построения классификации МВ

При обсуждении критериев разграничения типов МВ по степени конвенциональности мы исходили из нескольких эвристических принципов.

1) Принцип **верифицируемости**. Процедуры, применяемые для определения места конкретного МВ в классификационной схеме, должны быть проверяемыми и в минимальной степени опираться на компетенцию носителя языка.

2) Принцип **теоретической нейтральности**. Разные теории метафор — от архаичной теории замещения до теории концептуальной метафоры и когнитивного блендинга — отчетливо противостоят друг другу на уровне теоретических утверждений, но гораздо более согласованы на уровне эмпирического анализа. Если взять за основу конкретные приемы оценки метафор, предлагаемые разными теориями, то компромисс в части построения классификаций будет вполне достижимым. Полная теоретическая нейтральность вряд ли может считаться реалистичной целью, но уровень конфликтности теорий может быть снижен, если последовательно учитывать разные приемы анализа МВ.

3) Принцип **полноты**. Многие известные классификации ограничивают количество выделяемых классов тремя-четырьмя, причем разные классификации фокусируются на разных зонах шкалы конвенциональности (например, на отрезке языковых и узуальных метафор), рассматривая оставшуюся часть уже без деталей. Предлагаемая классификация, учитывая ее компромиссный и обобщающий характер, стремится к выделению максимального числа таксонов, а процесс оценки МВ сводится к алгоритмизированной процедуре.

Дифференциальные признаки степени конвенциональности МВ

В метафорологии последнего времени наметилась тенденция к разработке простых и прозрачных процедур описания метафор. Так, появились работы, нацеленные на выработку строгих критериев поиска и идентификации слов и выражений, использованных метафорически в текстах разных речевых регистров и жанров (Pragglejaz Group 2007), (Steen et al. 2010). В ряде исследований (Deignan 2005), (Goatly 1997), наряду с интуитивистскими признаками, обсуждается и ряд независимых от исследовательской интуиции критериев оценки метафор. Стимулом для этих попыток стала возможность использования лингвистических корпусов. Ниже предлагается список критериев и эвристических приемов идентификации, предлагаемых исследователями для разграничения разных типов МВ. В качестве иллюстраций использованы примеры из Национального корпуса русского языка (НКРЯ, ruscorpora.ru), однотомных толковых словарей русского языка, отдельные примеры креативных метафор привлечены из других источников.

1. Наличие или отсутствие словарной фиксации как исходного, так и метафорического выражения в толковом словаре. Словарь в этом случае выступает в качестве формализованного аналога семантической памяти среднего носителя. Этот критерий, во-первых, позволяет разграничивать инновационные, окказиональные, свежие, редкие метафоры, с одной стороны, и метафоры узуальные, уже ставшие фактами словаря, с другой. Кроме того, отсутствие словарной фиксации исходного значения может быть интерпретировано как индикатор финальной стадии жизни МВ в языке, его перехода в область этимологических метафор. В качестве разных типов фиксации можно рассматривать: 1) отражение значения в словарной статье в качестве единицы в нумерованном списке значений или в статусе подзначения, например *СЕТЬ 1. Приспособление, изделие из закрепленных на равных промежутках перекрещивающихся нитей, веревок, проволоки; 3. Система коммуникаций, расположенных на каком-нибудь пространстве* (Ожегов С. И., Шведова Н. Ю. Толковый словарь современного русского языка. 4-е изд., 2007. — далее СОШ); 2) наличие в словарной статье иллюстрации к метафорическому значению при отсутствии эксплицитного толкования: *Враг **не дремлет*** (СОШ) (= начеку, в состоянии готовности).

2. Возможность / невозможность замены метафорической единицы в контексте на неметафорический синоним. Этот способ проверки восходит к теории замещения, рассматривающей метафору как вторичный инструмент номинации. Невозможность естественного деметафоризирующего перифразирования является признаком стертой или мертвой метафоры (Black 1962): ср. *носик чайника, компьютерная мышка, ножка стула, глухой (звук), часы идут* vs. *продать своих друзей → предать своих друзей, дикая выходка → нелепая, странная выходка, острый ум → проницательный ум*. Этот признак позволяет

проводить границу между номинативными и фигуративными метафорами, являющимися добавочными средствами обозначения. Свойство служить средством «дополнительного наименования» (additional naming) входит в число главных эвристических критериев идентификации фигуративных выражений в (Dobrovol'skij, Piirainen 2005: 18–20).

3. Возможность или невозможность реализации метафорического значения в минимальном контексте, вне речевого окружения, содержащего лексические приметы области цели (Cruse 2004: 72), (Deignan 2005: 42–43). Например, лексема *жажда* 2 'сильное, страстное желание чего-н.' может реализовать это значение исключительно в составе конструкций с генитивом или зависимым инфинитивом типа *жажда славы*, *жажда знаний*, *жажда учиться*, а при употреблении лексемы в минимальном контексте типа *У него жажда* в нормальном случае метафорическое значение не актуализируется, и лексема интерпретируется только как 'он хочет пить'. Совсем иначе ведут себя лексемы *чистка 1* и *чистка 2*.

Пример 1. Метафорическая лексема *чистка 2* не нуждается в контекстуальной поддержке со стороны синтаксически зависимых компонентов и может актуализироваться как в разъясняющем контексте, насыщенном приметами области цели (*идеологическая чистка, чистка рядов партии),* так и вне его. Так, фраза *У нас опять была чистка* может быть интерпретирована и как 'у нас была уборка', и как 'у нас были увольнения с работы, исключения из организации или преследования на идеологических основаниях'. Проверка МВ по этому признаку существенно облегчается в случае использования корпуса, что показано в (Deignan 2005: 42–44). Обязательное наличие в синтаксическом контексте лексических примет области цели для реализации метафорического значения говорит о меньшей степени конвенциональности МВ по сравнению с контекстуально независимыми МВ, ср. (1):

(1) *Никакой чистки мы не проводили* [Немцов, Кремль и есть правительство (2003) // «Газета», 2003.06.20].

4. Метафоры, не фиксируемые словарями, естественно искать в корпусе. Нелексикографированные МВ могут быть редкими и более или менее равномерно распределяться во времени (пример 2), а могут концентрироваться в хронологически недавних контекстах, что свидетельствует об интенсивном вхождении в язык нового метафорического значения (пример 3).

Пример 2. Глагол *дремать 2* 'бездействовать, находиться в состоянии покоя' (ср. *вулкан дремлет, дремлющий вулкан*) не фиксируется в однотомных толковых словарях, в отличие от лексикографированного метафорического значения глагола *спать*, однако в НКРЯ обнаруживается не менее десятка примеров реализации *дремать 2* в текстах XIX–XXI века:

(2) *Нужен мониторинг — постоянное отслеживание тихих сигналов* ***дремлющего вулкана*** [Виталий. Вулкан по имени Эльбрус // Труд-7, 2007].

Пример 3. Выражение *(ломать) через колено* в метафорическом значении ‘решительно, радикально менять, возможно с разрушительными для объекта последствиями’ не отражено в СОШ, однако в основном и газетном подкорпусах НКРЯ зафиксировано не менее 16 и 30 вхождений соответственно, причем практически все они относятся к 1990–2000 гг. (только один пример датирован 1930 г.):

(3) *[Поправки] в очередной раз* ***через колено переломают*** *нефтянку* [Хайруллин. Новый бензиновый беспредел // «Московский комсомолец», 2003].

5. Если анализируемое МВ не зафиксировано ни в словаре, ни в корпусе, можно проверить его на соответствие концептуальным метафорам, определяющим строй того или иного языка. Такого рода разграничение предложено в работе (Kövecses 2002: 31–32). Лексически уникальное МВ может вполне укладываться в систему концептуального кросс-доменного сближения, поддержанного множеством других МВ (примеры 4–5 ниже), а может быть концептуально уникально и манифестировать индивидуальное совмещение двух областей опыта, не имеющее конвенциональных аналогов (пример 6).

Пример 4. В (4) используется не зафиксированное в словарях и не отраженное в НКРЯ выражение *рыночные колесики*:

(4) *Без [организаторских талантов] «труд мужика или ученого» не сцепляется* ***с рыночными колесиками****, сложная система не складывается!* [Радзиховский. Правда и революция // «Российская газета», 2010].

Это МВ реализует вполне стандартную для русского языка концептуальную механистическую метафору экономика — это механизм / транспортное средство (ср. узуальные *рыночные механизмы, локомотив экономики, двигатель экономики, тормоз экономического развития, экономические рычаги* и т. д.). Здесь можно говорить, пользуясь терминами фреймовой семантики, об индивидуальном заполнении слота в стандартном сближении фреймов.

Пример 5. МВ в (5)–(7) эксплуатируют целую совокупность близких концептуальных метафор страна / народ — это лошадь (домашний скот), глубоко укорененных в языке и родственных традиционному библейскому образу народа как стада (ср. стандартные *обуздать народ, держать народ в узде* и несколько менее стандартный *хомут крепостничества*):

(5) *Хотелось [Горбачеву]* ***ввести Россию в цивилизованное стойло****, да больно брыкастая она* [А. Яковлев. Омут памяти. (2001)];

(6) *Династия Романовых <…> **загнала свой народ в стойло крепостничества*** [Долуцкий. Дурная болезнь // «Век XX и мир», 1992];

(7) *[П]ришествие в клуб ЕС <…> **застоявшихся в советском стойле новобранцев** из Восточной Европы размыло влияние Франции в Евросоюзе* [Сантехник остался за дверью // РИА Новости, 2005].

Пример 6. Авторское МВ *разрушение слоя интеллектуального озона* из (8) не соответствует никакой готовой концептуальной метафоре (*образованные слои общества — это озоновый слой). Эта метафора находится на высшей ступени неконвенциональности, реализуя непрогнозируемую проекцию из физики атмосферы в область социальной реальности.

(8) *Откуда серый туман скуки и падение драйва, **разрушение** ноосферы, **слоя «интеллектуального озона»** над страной?* [Л. Радзиховский, «Российская газета», 2010].

Классификация типов МВ по степени конвенциональности

На основе этих признаков можно построить классификацию типов метафор по степени конвенциональности, отражающую шкалу постепенных переходов от наиболее старых, стертых метафор до метафор креативных. Предлагаемая ниже классификация имеет априорный характер, в будущем предполагается ее использование при анализе метафорического строя финских политических текстов и их переводов на русский язык.

а) **Этимологическая** метафора. В толковых словарях фиксируется только вторичное значение таких МВ. Например, лексема *иго* в современных русских словарях толкуется только как ‘угнетающая, порабощающая сила’ (СОШ), однако в истории языка это слово развивает вышеупомянутые современные значения на основе утраченного ‘ярмо, хомут’.

б) **Мертвая** метафора. Словари фиксируют как первичное, так и вторичное, метафорическое значения слова, при этом метафорическое значение выступает в качестве главного средства номинации определенных объектов, свойств, действий, состояний и не может быть оформлено другими (неметафорическими) лексемами: *альпинистские **кошки**, часы **идут, высокий** голос, **низкое** давление.* В интуитивистских терминах такие метафоры часто называются **безóбразными**.

в) **Спящая** метафора. Толковые словари так же, как и в предыдущем случае, фиксируют как первичное, так и метафорическое значения слова, однако переносное значение слова не является единственным средством номинации и может быть естественным образом заменено на неметафорическое: *идет снег→падает снег, продать своих друзей→предать своих друзей*. Дифференциальным признаком этого типа является высокая степень

независимости метафорического значения от лексического и конструктивного окружения. Метафора успешно может быть реализована в контексте, не насыщенном лексическими элементами области цели. См. также пример *чистка* выше.

г) **Конвенциональная** метафора. От спящей метафоры этот тип отличается лексико-синтаксическими связями МВ с элементами контекста, однозначно указывающими на область цели. Вне такого контекста реализация метафорического значения затруднена. Так, лексема *поток* может реализовать метафорическое значение ‘движущаяся масса чего-н., кого-н.’ только в контексте зависимых имен, находящихся в лексическом контрасте с исходным значением слова: *людской поток, поток мыслей, поток слов, поток света*.

Ниже описываются разновидности нелексикографированных метафор.

д) **Редкая (речевая)** метафора. В этот класс входят МВ, неоднократно, хотя и нечасто фиксируемые в корпусе. Это неуникальные, хронологически и жанрово распределенные МВ. Как правило, они представляют собой прогнозируемое развитие уже имеющихся в языке концептуальных метафор, например, лексема *инфляция* часто используется в неэкономическом метафорическом значении ‘обесценивание, разрушение ценности или значения чего-н.’, которое лишено лексикографической фиксации. В НКРЯ, однако, обнаруживается немало примеров реализации этого значения, реализующегося обычно в составе генитивной конструкции: *инфляция понятия, инфляция слова, инфляция доверия, инфляция искусства, инфляция ценностей*. Это значение слова *инфляция* вполне укладывается в распространенную концептуальную метафору значения / ценности / продукты ментальной деятельности человека (убеждения, мысли, идеи и т. п.) — это объекты, имеющие цену: ср. *его слову грош цена, он дорого заплатил за свои убеждения, ценная мысль, дешевая риторика*, *торговать убеждениями*.

е) **Свежая** метафора. МВ этого типа, разделяя с речевыми метафорами такой признак, как неоднократная корпусная фиксация, отличаются от последних тем, что характеризуют последний этап существования языка и постепенно наращивают частотность. Они могут войти в язык и попасть в словарь, однако для этого должно накопиться репрезентативное количество их употреблений (ср. пример *ломать через колено* выше). Свежие МВ могут как соответствовать уже манифестированным в языке концептуальным метафорам, так и устанавливать новое концептуальное соответствие между доменами, увеличивая спектр кросс-доменных сближений. К примеру, на рубеже 1980–1990 гг. в публицистических и художественных текстах стали появляться лексические манифестации новой органистической метафоры человеческое сообщество — это грибница, которая порождает образ структурированной человеческой группы, обладающей отсутствием централизованной структуры и свойствами способности к самовоспроизведению, естественности, тенденции к самоорганизации, невидимости. МВ с лексемой *грибница* в значении ‘общество, социальная структура’ стабильно, хотя и нечасто используются на протяжении 30 лет.

(9) *Когда-то у нас в стране была подлинная* ***грибница интеллигентных людей****. Эта* ***грибница*** *плодила <…> настоящих, благородных людей, для которых важнее всего был голос их совести. Эту* ***грибницу*** *и уничтожили в несколько туров* [В. Дудинцев. Цвет наших одежд // «Литературная газета», 1988];

(10) *В его рассказах автор не отличается от героев, потому что все* ***люди*** *для Довлатова были* ***из одной грибницы*** [А. Генис. Довлатов и окрестности, 1998]

(11) *Нельзя до поры тормошить разросшуюся в российской элите* ***ельцинскую «грибницу»*** [Раз — наперсток // «Санкт-Петербургские ведомости», 2003].

ж) **Инновационная** метафора. Это лексически уникальные, крайне редкие в корпусе, но концептуально предсказуемые метафоры, представляющие собой результат окказионального заполнения целевого слота при общей когнитивной поддержке системы, при наличии манифестированных в узусе фактов сближения доменных областей. К примеру, одна из стандартных метафорических концептуализаций состояния сна — это вместилище, заполненное жидкостью: ср. *глубокий сон, поверхностный сон, погрузиться в сон, сон улетучился*. Эта концептуальная метафора имеет лексические лакуны, которые легко заполняются в индивидуальном дискурсе:

(12) *Я чуть* ***вынырнул*** *из сна* [В. Попов, Будни гарема, 1994];

(13) *Тихо* ***стекает*** *с тела сон* [Т. Толстая, Ночь, 1983].

з) **Креативная** метафора. МВ этого типа уникальны и концептуально, и лексически: они возникают в результате не имеющего аналогов сближения доменных областей, не поддержанного стоящей за языком концептуальной системой (ср. *слой интеллектуального озона*). В иллюстрации (14) в оригинальных выражениях манифестируется слабо укорененное в русском политическом дискурсе концептуальное сближение политические процессы — это приготовление еды, единичные реализации которого встречаются в общественно-политической речи (*кремлевские повара, политический рецепт*).

(14) *Из* ***полусырого бульона ельцинского времени*** *почти уже* ***сварился авторитарно-олигархический супчик*** [Прибыловский, Неприкосновенный запас, 2004, № 38].

Представленная классификация может быть представлена в виде алгоритма.

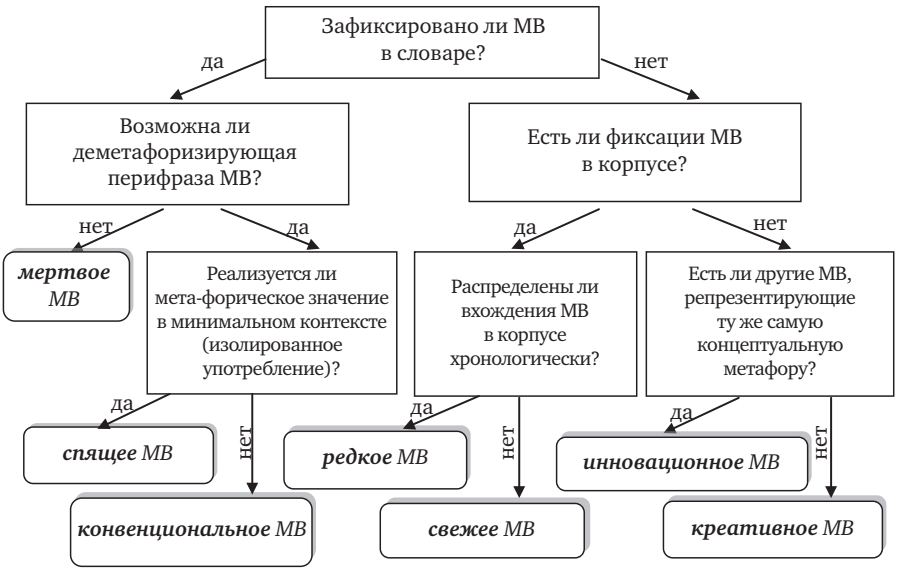


Схема 1. Алгоритм оценки метафорических выражений (МВ) по степени креативности/конвенциональности

References

1. *Baranov A. N.* (2003) Political metaphors in political texts: the possibilities of linguistic monitoring. [Politicheskaia metaforika politicheskogo teksta: vozmozhnosti lingvisticheskogo monitoringa]. Iazyk SMI kak obiekt mezhdistsiplinarnogo issledovaniia. Moscow. Available at: ttp://evartist.narod.ru/text12/09.htm.
2. *Bally Ch.* (1951), Traité de stylistique française. Geneve.
3. *Black M.* (1962), Models and metaphors: studies in language and philosophy. Ithaca, N.Y.
4. *Cruse A. D.* (2004), Meaning in Language. An Introduction to Semantics and Pragmatics. Oxford.
5. *De Landtsheer Ch.* (1988), The political rhetoric of a United Europe, in Politically speaking: a worldwide examination of language used in the public sphere. New York, pp. 129–146.
6. *Deignan A.* (2005), Metaphor and Corpus Linguistics. Amsterdam.
7. *Dobrovol'skij D., Piirainen E.* (2005), Figurative Language: Cross-Cultural and Cross-linguistic Perspectives. Amsterdam.
8. *Goatly A.* (1997), The Language of Metaphor. London, New York.
9. *Kövecses Z.* (2002), Metaphor. A Practical Introduction. Oxford, 2002.
10. *Pragglejaz Group* (2007), MIP: A method for identifying metaphorically used words in discourse, Metaphor and Symbol, 2007, 22 (1), pp. 1–39.

11. *Steen G. J., Dorst A. G., Herrman J. B., Kaal A. A., Krennmayr T., Pasma T.* (2010), A Method for Linguistic Metaphor Identification. Amsterdam: John Benjamins, 2010.
12. *Vertessen D., De Landtsheer C.* (2005), "Metaphorical election style? Patterns of symbolic language use in Belgian politics" available at: eis.bris.ac.uk/~potfc/Granada/Papers/Vertessen.pdf.

BELARUSIAN AND RUSSIAN LINGUISTIC PROCESSING MODULES FOR THE SYSTEM NOOJ AS APPLIED TO TEXT-TO-SPEECH SYNTHESIS

Hetsevich Yu. S. (Yury.Hetsevich@gmail.com)
United Institute of Informatics Problems of the National Academy of Sciences of Belarus (UIIP NAS Belarus), Minsk, Belarus

Hetsevich S. A. (Novaeimya@gmail.com)
UIIP NAS Belarus, Minsk, Belarus

Lobanov B. M. (Lobanov@newman.bas-net.by)
UIIP NAS Belarus, Minsk, Belarus

This paper describes the program NooJ, which provides a platform for the construction of modules for resolving linguistic problems in the area of text-to-speech synthesis. Belarusian and Russian are chosen as target languages. Basic and comprehensive electronic grammatical dictionaries for NooJ are described. We present the entire algorithm of converting dictionaries of the two languages into the form acceptable in NooJ, which retains all lexical, grammatical and accentual information. The dictionaries developed for NooJ help solve the problems of annotating words with lexical and grammatical categories and syllabic accents, as well as the problems of searching texts for a definite sequence of words according to their grammar and word forms. The grammars developed for NooJ are notable for their clarity and the speed at which they can be built. The structure and algorithm of the morphological grammar working on the text are given, including the localization of words written in Cyrillic and Roman letters. The structure and algorithm of the work of two syntactic grammars on a text are described: one aimed at searching for phonetic words and the other on searching for syntagms with a particular number of phonetic words. NooJ output produced after dictionaries or grammars have been applied to text are exported to text format Future work includes the completion of the transfer pf accentual information into Belarusian and Russian NooJ dictionaries,and construction of grammars for identifying accentual units in syntagms and grammars aimed at learning rhythmic structures of texts.

Key words: text-to-speech synthesis, NooJ, Belarusian language, Russian language, linguistic modules, dictionary, morphological grammar, syntactic grammar, text annotation, syntagmation, phonetic words

1. Introduction

The linguistic program NooJ (Silberztein 2003) contains numerous dictionaries, grammars and processed corpuses. The internal linguistic algorithms in NooJ permit the processing of separate files and also collections of files (corpora). For text processing, the following basic operations are developed: selection of tokens, bigrams, and unknown words, unambiguous and ambiguous words.

NooJ helps to create large lexical, morphological and syntactic grammars, which can be written both in text view and through an easy-to-use visual editor, greatly simplifying the process. Dictionaries and grammars can be applied to texts to find complex lexical, morphological and syntactic expressions. NooJ users are able to develop grammars for identifying semantic units such as proper names, dates, time, financial expressions[1], etc., in a large text.

In 2011 the authors had the opportunity to develop the first version of Belarusian and Russian NooJ modules. The discussion about the results took place at the international conference of the NooJ community in Dubrovnik, Croatia (Hetsevich 2012; Hetsevich 2012). This article will present an improved way of building Belarusian and Russian dictionaries for NooJ. Particularly for use in text-to-speech synthesis, more accentual information for words will be added from the basic electronic dictionaries. Grammars important for linguistic text processing in text-to-speech synthesis will also be examined.

2. Basic Belarusian and Russian electronic dictionaries for text-to-speech synthesis system Multiphone

The bilingual text-to-speech synthesis system Multiphone uses specially-prepared electronic dictionaries for Belarusian and Russian languages (Lobanov 2008). For the basic electronic dictionaries, a system of codifying lexical and grammatical categories was developed. The initial Belarusian electronic dictionary (BED) was based on the printed dictionary Biryla (Biryla 1987, Sovpel 2006). The Russian electronic dictionary (RED) was built in SSRL using the printed Zaliznyak dictionary (Zaliznyak 1980).

New words from modern texts on literature, science and technology were added to the basic electronic dictionaries. To further facilitate the process, one of the authors has developed an application program, Expert Dictionary Editor (EDE) (Hetsevich 2011). EDE makes it possible to process texts, find new words for the dictionary, and allows the user to mark syllabic accents, lexical and grammatical categories (LGC) and priorities therein. Saved words (including their homographs) can be displayed together with all of the selected information (Fig. 1).

[1] http://www.nooj4nlp.net/pages/introduction.html

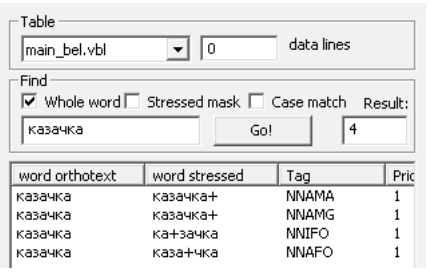


Fig. 1. The program EDE makes it possible to view, add and delete words in electronic dictionaries

Quantitative characteristics of the dictionaries (table 1) show that each dictionary has 14 lexical categories (or simply categories) of words. Category refers to part of speech (noun, verb, etc.), as well as several word groups (for instance, parenthesis, predicate). Each category is marked with the amount of corresponding word forms and tags.

Table 1. Content of Belarusian and Russian electronic dictionaries

	BED		**RED**	
Categories	**Word forms**	**Tags**	**Word forms**	**Tags**
Adjective	872 943	85	504 728	31
Adverb	6524	8	1366	33
Conjunction	45	2	563	49
Gerund	32 278	4	70 215	8
Interjection	8	1	184	1
Noun	592 245	244	540 684	168
Numeral	3571	96	1473	92
Parenthesis	45	1	65	1
Participle	189 609	116	1 897 033	246
Particle	57	3	348	75
Predicate	60	1	266	1
Preposition	121	15	277	32
Pronoun	1214	270	1064	307
Verb	419 734	68	429 587	78
Sum	**2.118.454**	**914**	**3.447.853**	**1122**
Sum(KB)	**41 267,2**	**50,5**	**92 774,4**	**80,6**

The total number of entries in the Belarusian electronic dictionary is more than 2,1 million word forms, and in Russian 3,4 million word forms. The total number of tags in each of the dictionaries is 914 and 1122 respectively. The total size of the files with word forms and tags for BED is 40,4 MB, and for RED 90,7 MB.

The dictionaries effectively manage the problem of identifying syllabic accent and LGC in a wide range of thematic texts for text-to-speech synthesis. Further, we shall try to realize the conversion of the aforementioned electronic dictionaries into NooJ so that they can be used by linguists more in processing texts in Belarusian and Russian.

3. Building Belarusian and Russian dictionaries for NooJ

For converting Belarusian and Russian electronic dictionaries into NooJ, various algorithms of conversion have been developed in Perl (Fig. 2) . The algorithms receive BED and RED in input, process them and put out dictionary files in NooJ according to the rules of its construction (Silberztein 2003). In place of XX is a language identifier: "be" for Belarusian and "ru" for Russian.

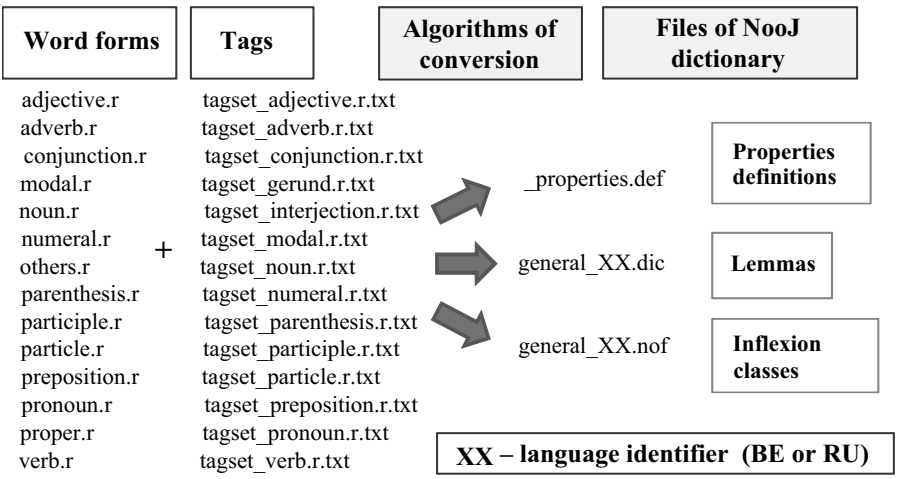


Fig. 2. Converting Belarusian and Russian electronic dictionaries into NooJ

Let us examine the steps of processing electronic dictionaries with developed algorithms.

Firstly the file with lexical property definitions entitled ***_properties.def*** is formed. For this all the tags and their decodings from the tag sets (***tagset_*.r.txt***) are read, but the only ones kept are those which are really used for marking LGC of words in the word file of a specific category (****.r***). For each property of the category all possible variants of their values are collected from the respective columns with tag decoding. They are written with "+" sign in the following format:

CATEGORY_propertyName = propertyVALUE1+propertyVALUE2 +…;

For example, fig. 4 presents excerpts of obtained files with property values for the category Noun.

Obtained from BED	**Obtained from RED**
… NOUN_Animation = Animate + Inanimate; NOUN_Case = Accusative + Dative + Genitive + Instrumental + Nominative + Prepositional; NOUN_Form = SubstanceAdjective; NOUN_Gender = Feminine + Masculine + Neuter; NOUN_Meaning = Common + Proper; NOUN_Flexion = ю + яў_аў; NOUN_Number = Plural; …	… NOUN_Animation = Animate + Inanimate; NOUN_Case = Accusative + Dative + Genitive + Instrumental + Nominative + Prepositional; NOUN_Gender = Feminine + Masculine + Neuter; NOUN_Number = Plural + Singular; NOUN_ProperCommon = Common; …

Fig. 4. Excerpts of the files with properties for the category Noun

Then the algorithm forms a file with word lemmas entitled general_XX.dic. For each paradigm of a specific category, lemma and inflection class with tags are formed. A lemma is defined as the initial form of a word. The inflection class for each lemma is of the form "inflection1_tag1, inflection2_tag2, …, inflectionN_tagN". Each inflection is obtained by deleting the minimal part of a word that remains invariable in all word forms of a word paradigm from the word form. Every obtained inflection class is checked according to the following rule:

Is there already such an inflection class with tags in the file ***general_XX.nof*** or not?

1. If not, the inflection class is registered in the file ***general_XX.nof*** with the resulting title being identical to the lemma. If there is such a class title, it is completed with a unique number. Couples of flexions and tags alternate with the sign "+". Commands to delete part of a lemma (number of characters) are inserted before the flexions, after which the NooJ parser adds flexions of its inflection class to the lemma. Decoding is taken from the file with decoding for each tag, the category title is deleted and the sign "+" is added between property values. The end of the inflection class is denoted by the sign ";".
2. If there is such an inflection class, it is already being used for some word, and a name has already been created for it.

Figure 5 contains excerpts of files with inflection classes for the category Noun.

Obtained from BED	**АБАВЯЗАК** = <B2>ак/Accusative+Common+Inanimate+Masculine+ <B2>ак/ Common+Inanimate+Masculine+Nominative+ <B2>кі/Accusative +Common+Inanimate+Masculine+Plural+ <B2>кі/Common+Inan imate+Masculine+Nominative+Plural...;…
Obtained from RED	**АБСОЛЮТНОСТЬ** = <B1>ей/Common+Feminine+Genitive+Inan imate+Plural + <B1>и/Accusative+Common+Feminine+Inanimate+Plural + <B1>и/Common+Feminine+Inanimate+Nominative+Plural + <B1>и/Common+Dative+Feminine+Inanimate+Singular + <B1>и/Common+Feminine+Genitive+Inanimate+Singular...; …

Fig. 5. Excerpts of files with inflection classes for the category Noun

In order to mark accent, the algorithm collects all accent positions from each word worm of a lemma. If all accent positions are equal to the number A, a marker with the constant "+sA" is formed. If there is even one different accent position in any word worm of lemma, the marker of various accent in the paradigm is formed — "+sN".

The lemma is registered in the file ***general_XX.dic*** with the mark of the category, plus the command "+FLX=" and then the title of the corresponding inflection class in the following format:

LEMMA, CATEGORY + FLX = TITLE_OF_INFLECTION_CLASS + ACCENT_MARKER.

Figure 6 presents the resulting files with lemmas, marked categories, inflection classes and accents for BED and RED.

Obtained from BED	… абавязак,NOUN+FLX=АБАВЯЗАК+s5 адгалосак,NOUN+FLX=АБАВЯЗАК+s6 мама,NOUN+FLX=АБАТЫСА+s2 манастыр,NOUN+FLX=АБРУЧ+sN …
Obtained from RED	… абстрактность,NOUN+FLX=АБСОЛЮТНОСТЬ+s6 аварийность,NOUN+FLX=АБСОЛЮТНОСТЬ+s5 адоптировать,VERB+FLX=АБЛАКТИРОВАТЬ+s6 быстр,ADJECTIVE+FLX=БОДР+sN …

Fig. 6. Excerpts of files with lemmas

In the beginning of every resulting file of the NooJ dictionary, a corresponding heading with the mark of the NooJ version (V3), authors and other necessary information is added to save the correct format of the dictionary files.

Thereby, the resulting files ***_properties.def***, ***general_XX.dic*** and ***general_XX.nof*** correspond to the rules of construction for files in the NooJ dictionaries.

Quantitative characteristics of the resulting Belarusian (BN) and Russian (RN) NooJ dictionaries are presented in Table 2. There are nearly 137 thousand lemmas in BN 123 565 (88 %) of which are lemmas with precise accent position. There are nearly 214 thousand lemmas in RN, 199 638 (94 %) of which are lemmas with precise accent position. The total number of inflection classes in BN is 3137 and 1851 in RN. The number of property values in BN and in RN is 61 and 56 respectively. These numbers of property values are almost identical. This can be explained by the fact that Belarusian and Russian are kindred languages, having similar grammatical and categorical structure.

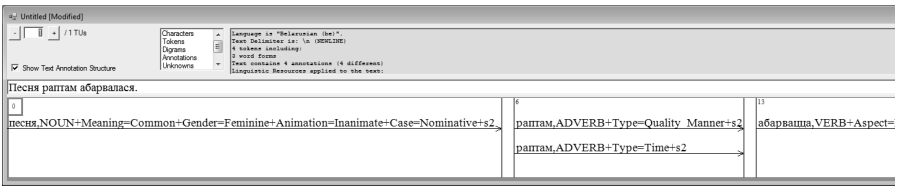
Table 2. Quantitative characteristics of Belarusian and Russian electronic dictionaries for NooJ

	Lemmas		**Inflection classes**		**Property values**	
Categories	**BE**	**RU**	**BE**	**RU**	**BE**	**RU**
Adjective	32 350	31 583	33	61	8	5
Adverb	6524	1366	8	30	2	4
Conjunction	45	563	2	49	1	4
Gerund	1836	32 977	282	160	2	4
Interjection	8	184	1	1	0	0
Noun	53 054	44 396	1427	309	8	5
Numeral	169	99	38	27	8	5
Parenthesis	45	65	1	1	1	0
Participle	12 293	68 006	13	57	8	9
Particle	57	348	3	73	2	5
Predicate	60	266	1	1	1	0
Preposition	105	277	17	27	2	4
Pronoun	66	1064	56	307	9	5
Verb	30 718	33 127	1255	748	9	6
Sum	**137 330**	**214 321**	**3137**	**1851**	**61**	**56**
Sum (KB)	**6064**	**10 446**	**1809**	**1113**	**5**	**5**

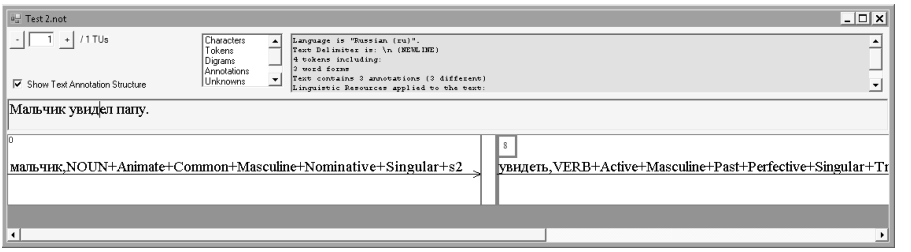
The total size of BN (the file with lemmas, inflection classes and property values) is 7,7 MB, and that of RN is 11,3 MB. If these files are processed with the NooJ parser to create a dictionary in one file, the result will be a file of the type ***general_XX.nod***. For Belarusian it will take up 8,2 MB, and for Russian 5 MB.

4. Linguistic processing of Belarusian and Russian texts with developed dictionaries for text-to-speech synthesis

When Belarusian and Russian dictionaries for modules of the program NooJ are developed, they can be used for text annotating. For instance, for each known word, all variants of accents and grammatical categories present in the dictionary are marked (Fig. 7).



(a)



(b)

Fig. 7. Examples of annotating words with accents and grammatical categories for Belarusian (a) and Russian (b) languages

It is possible to receive an XML-export of the annotated sentence (Figure 8). For each word, the export contains its lemma, category, inflection class name, expanded grammatical marks and accent marker. As such, this information can be processed further using the linguistic processor of the text-to-speech synthesizer.

BN	<LU LEMMA="**песня**" CAT="NOUN" FLX="ПЕСНЯ" Meaning="Common" Gender="Feminine" Animation="Inanimate" Case="Nominative" TYPE="s2">**Песня**</LU> <LU LEMMA="**раптам**" CAT="ADVERB" FLX="АБАВЯЗАЦЕЛЬНА" Type="Quality_Manner" TYPE="s2"> <LU LEMMA="**раптам**" CAT="ADVERB" FLX="АДВЕКУ" Type="Time" TYPE="s2">**раптам**</LU></LU> <LU LEMMA="**абарвацца**" CAT="VERB" FLX="АБАРВАЦЦА" Aspect="Perfective" TYPE="sN">**абарвалася**</LU>.

RN	<LU LEMMA="**мальчик**" CAT="NOUN" FLX="АБРЕК" TYPE="Animate" TYPE="Common" TYPE="Masculine" TYPE="Nominative" TYPE="Singular" TYPE="s2">**Мальчик**</LU> <LU LEMMA="**увидеть**" CAT="VERB" FLX="ВОЗНЕНАВИДЕТЬ" TYPE="Active" TYPE="Masculine" TYPE="Past" TYPE="Perfective" TYPE="Singular" TYPE="Transitive" TYPE="s3">**увидел**</LU> <LU LEMMA="**папа**" CAT="NOUN" FLX="БОНЗА" TYPE="Accusative" TYPE="Animate" TYPE="Common" TYPE="Masculine" TYPE="Singular" TYPE="s2">**папу**</LU>.

Fig. 8. Examples of XML-export of the data annotated from NooJ

The resulting dictionaries have been tested in NooJ using two types of queries for Belarusian and Russian texts:

- To find words and their contexts so that they correspond to the given categories with indicated property values (Figure 6);
- To find words and their contexts so that they correspond to word forms of the given paradigm (Figure 7).

(The paradigm is set through any word form of this paradigm).

query for BN: <NOUN><NOUN+Genetive>

Before	Seq.	After
трапятаў	агеньчык каганца	. Каля
ліловым	адценнем пліткі	жалезняку
ведаць	асновы свету	, што
старажытныя	белыя званіцы	. А
песню.	Бог салдата	свайго

query for RN: <VERB><ADVERB>

Before	Seq.	After
человека, который	вставал очень	рано
части, и	было немножко	страшно
позднего вечера	бродил где-нибудь	. Однажды
кончились; я	прошел мимо	белого
слушать ее	было интересно	, хотя

Fig. 9. Locating words and their contexts by the given categories and property values

query for BN: <горкі>

Before	Seq.	After
На	горцы	лясок
Каля	горкі	сям
Чаму	горкія	асіны

query for RN: <жить>

Before	Seq.	After
помнил,	жила	моя
редактор.	Живет	в
он	живет	. Выяснилось

Fig. 10. Locating words and their contexts, where words correspond to word forms of the given paradigm

As a result, it is possible to study the forming correlations between words in sentences according to grammatical features of words.

In summary, the Belarusian and Russian dictionaries for NooJ help to solve basic problems with the linguistic aspect of the text-to-speech synthesizer and to study the

correlations between words in texts for the building of syntactical grammars in order to define phonetic words and accentual units in text.

5. Linguistic processing of Belarusian and Russian texts with morphological and syntactical grammars for text-to-speech synthesis

NooJ makes it possible to develop graphic grammars for solving common problems with processing specific words and their sequences in text-to-speech synthesis. This program does not require knowledge of a programming language; it has a special visual editor for building grammars.

Input texts can contain many words unknown to dictionaries, which causes the words to be read incorrectly in the text-to-speech synthesis system. This can be the case with words formed from mixed alphabets (Cyrillic and Roman, for example). And the most difficult is the case when two similar symbols have different codes for Belarusian/Russian and English languages (/i/, /i/), (/c/, /c/). For the localization of such words, a special morphological grammar has been developed with the visual editor NooJ (Figure 11). This grammar performs the following actions:

- marks foreign words with the category EN (searches words containing all letters from the set ENGLISH_LETTERS and marks them);
- marks words with mixed alphabets with the category MIXED and marks Roman symbols with brackets (Figure 12);

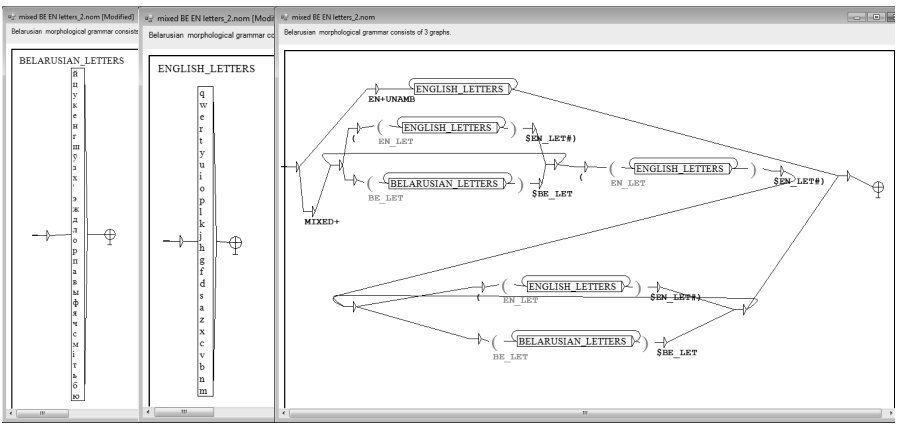
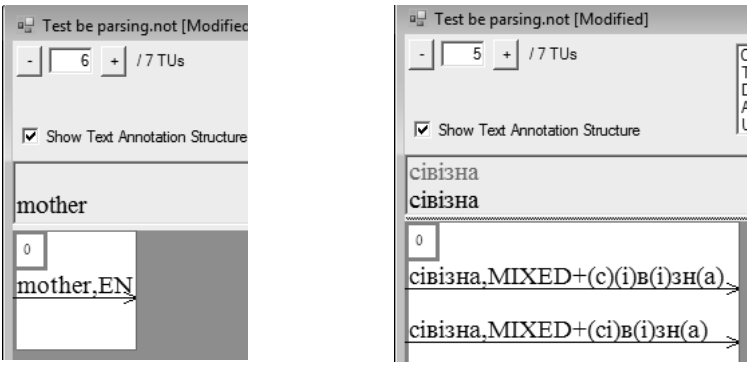


Fig. 11. Example of building morphological grammar for finding words written in mixed alphabets

For convenience we shall understand ***complex phonetic words*** as words formed from two or more words, and ***phonetic words*** as those formed from only one word.

In order to mark complex phonetic words, a syntactical grammar has been developed (Figure 13). The grammar works in a way that left and right context of a word is checked, and according to its meaning, the word sequence can be marked as a complex phonetic word in the following ways (Figure 14). If there are prepositions (one or more) to the right of the word, such a word sequence is marked as a complex phonetic word (PHONETIC_WORD) formed with prepositions (+PRE), for example, /**у кавалкі**/, /**на куски**/. If there are particles (one or more) to the left of the word, such a word sequence is marked as a complex phonetic word (PHONETIC_WORD) formed with particles (+PAR), for example, /**што ж**/, /**что же**/. If the word has both prepositions (one or more) and particles (one or more) as right and left context, such a word sequence is marked as a complex phonetic word (PHONETIC_WORD) formed both with prepositions and particles (+PRE+PAR), for example, /**за што ж**/, /**за что же**/.



(a) (b)

Fig. 12. Examples of work of the morpholgical grammar for finding words written in Roman (a) and mixed (b) alphabets

NooJ makes it possible to import a grammar into another one and to use sub-grammars in grammar. In order to define possible syntagmas containing four phonetic words, a corresponding grammar has been developed (Figure 15). It marks word sequence with the category **SYNTAGMA+4**, if four phonetic words follow each other (Figure 15, a). In order to identify four phonetic words (including complex words), a sub-grammar **WF_or_PHW** is invoked. This sub-grammar is employed if a simple word is found (**WF**) or if a complex phonetic word is found (**phonetic_word**) according to the sub-grammar on the Figure 15, b. An example of work of the grammar on a specific text is shown in Figure 16.

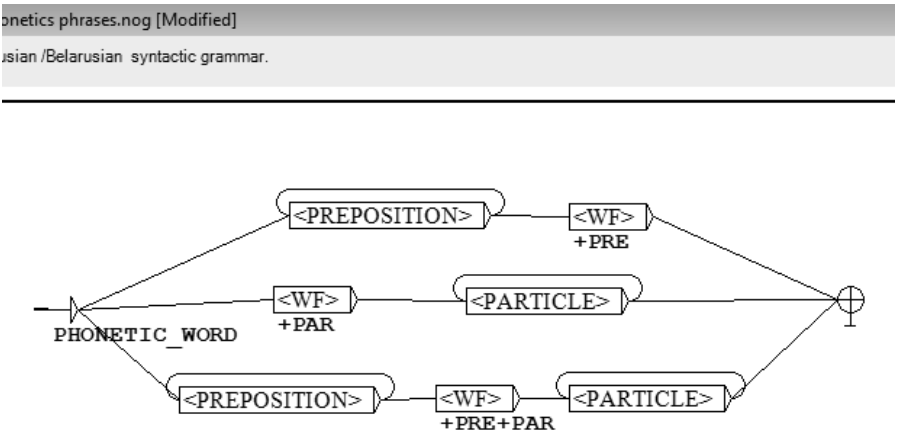
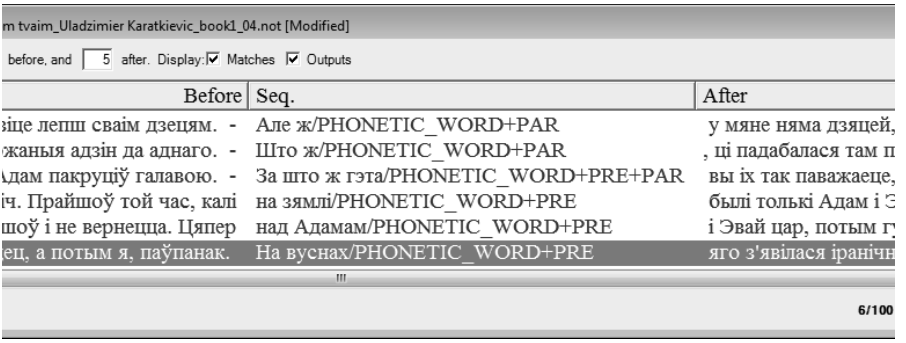
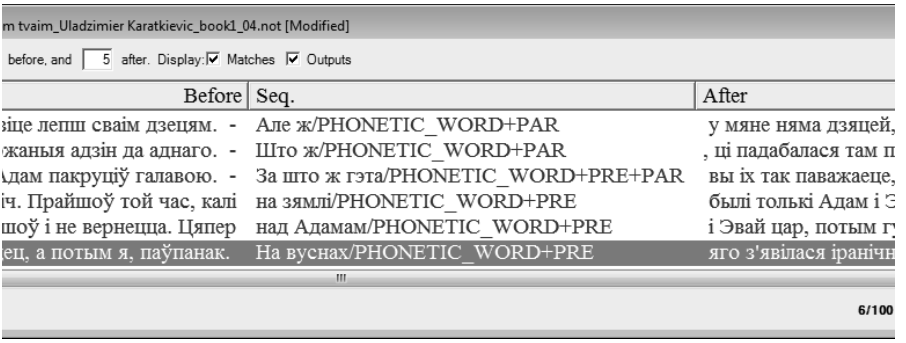


Fig. 13. Syntactical grammar for defining complex phonetic words

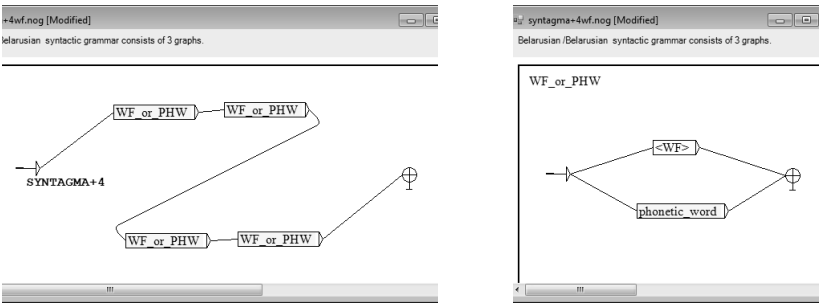


(a)



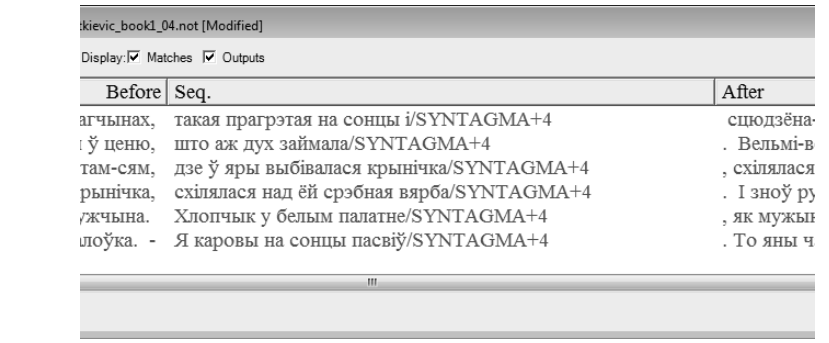
(b)

Fig. 14. Results of work of the syntactical grammar for defining phonetic words in texts in Belarusian (a) and Russian (b) languages

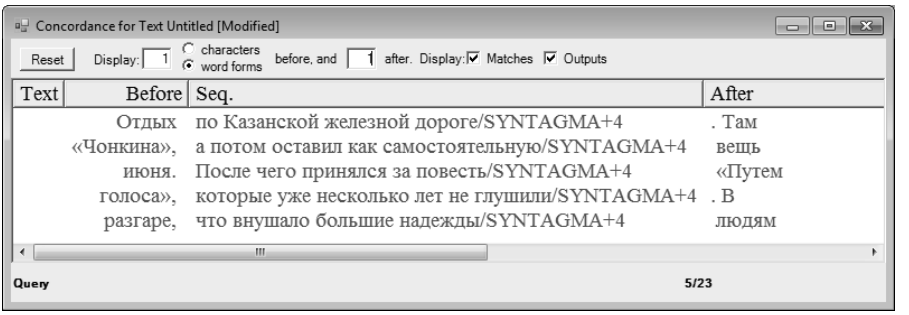


(a) (b)

Fig. 15. Syntactical grammar (a) for separating possible syntagmas into four phonetic words according to the sub-grammar (b)



(a)



(b)

Fig. 16. Examples of work of the syntactical grammar for separating possible syntagmas into four phonetic words for Belarusian (a) and Russian (b) texts

It should be noted that a grammar for separating syntagmas into four phonetic words can be extended to any number of phonetic words. This research gives an example of building grammars according the most frequent number of phonetic words in a syntagma (Lobanov, 2011).

Results of the work of the syntactical grammar on a text can be exported into a text file marked by the beginning of the given syntagma and its length. Such results can be processed with another grammar in order to form accentual units.

Conclusion

This article presents qualitative features of Belarusian and Russian electronic dictionaries which are used in the text-to-speech synthesis system Multiphone. Algorithms for converting basic dictionaries for Belarusian and Russian modules of linguistic resources of the program NooJ have been developed.

The total number of lemmas for the Belarusian dictionary is nearly 137 thousand, and for the Russian approximately 214 thousand. Information about precise accent position in lemmas' paradigms has been transferred to the Belarusian dictionary at 88% and to the Russian — at 94%. The resulting dictionaries are able to annotate words with lexical and grammatical categories and accents and also to help with studying the syntax of sentences.

NooJ tools for building visual morphological and syntactical grammars allow linguists to structure quickly and graphically important linguistic algorithms for text-to-speech synthesis: searching for words written in Cyrillic and Roman letters, searching for phonetic words, searching for syntagmas with a specific number of phonetic words.

The developed Belarusian and Russian NooJ modules will be improved. In addition, all accentual information will be transferred to dictionaries; syntactical grammars for defining accentual units in syntagmas will be developed; and the problems of learning rhythmic structure of texts will be solved.

The modules for dictionaries in NooJ can be also used by linguists and philologists in order to learn Belarusian or Russian languages.

Acknowledgments

Many thanks to the linguist from USA Adam Morrison for his help in revising this paper.

References

1. *Biryla M. V.* (1987), Slounik belaruskaj movy: Arfagrafija. Arfaepija. Aktsentuatsyja. Slovazmjanenne [Belarusian dictionary: Orthography. Orthoepy. Accentuation. Inflection]. Minsk: BelSE
2. *Hetsevich Y., Hetsevich S.* Overview of Belarusian and Russian dictionaries and their adaptation for NooJ, Selected Papers from the NooJ 2011 International Conference "Automatic Processing of Various Levels of Linguistic Phenomena", Dubrovnik, Croatia. Newcastle: Cambridge Scholars Publishing, 2012, pp. 29–40.

3. *Hetsevich Y., Hetsevich S., Lobanov B., Yakubovich Ya.* (2012) “Belarusian module for NooJ”. Available at: http://www.nooj4nlp.net/pages/belarusian.html.
4. *Hetsevich Y. S., Vyaltsev V. N.* Editing and completion system for the dictionaries of audio interface of question-and-answer system for Belarusian and Russian [Sistema redaktirovanija I popolnenija slovarej rechevogo interfejsa voprosno-otvetnoj sistemy dlja belaruskogo i russkogo jazykov]. Open Semantic Technologies for Intelligent Systems: Proceedings of the International Scientific and Technical Conference. Minsk, 2011, pp. 413–424.
5. *Lobanov B. M., Hetsevich Y. S.* Statistical Characteristics Of Sentence Syntagmatic Segmentation In Applying To Expressive Text-To-Speech Synthesis [Statisticheskie harakteristiki sintagmaticheskogo chlenenija predlozhenija v prilozhenii k sintezu vyrazitel'noj rechi po tekstu], Komp’iuternaja Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog 2011” [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2011”]. Moscow, 2011, pp. 434–447.
6. *Lobanov B. M., Tsirulnik L. I.* (2008), Kompjuternyj sintez i klonirovanie rechi [Computer synthesis and cloning of speech], Belaruskaja Navuka, Minsk.
7. *Silberztein, M.* (2003), “NooJ Manual”. Available at: www.nooj4nlp.net
8. *Sovpel, I. V.* Computer corpus of Belarusian language and its applications [Kompjuternyj fond belorusskogo jazyka i ego prilozhenija]. Materialy 3 mezhdunarodnoj konferentsii “Informatsionnye sistemy i tehnologii” IST’2006, chast’ 2 [Proceedings of the 3 International Conference “Information systems and technologies” IST’2006, part 2]. Minsk, 2006, pp. 71–76.
9. *Zaliznjak, A. A.* (1980), Grammaticheskij slovar’ russkogo jazyka: Slovoizmenenie. Ok. 100 000 slov [Russian grammatical dictionary: Inflection. Near 100 000 words], Stereotip, Moscow.

СЛОВАРЬ БЫТОВОЙ ТЕРМИНОЛОГИИ: НОВЫЕ ПРОБЛЕМЫ И НОВЫЕ МЕТОДЫ[1]

Иомдин Б. Л. (iomdin@ruslang.ru),
Лопухина А. А. (nastya-merk@yandex.ru)
Институт русского языка им. В. В. Виноградова РАН

Пиперски А. Ч. (apiperski@gmail.com)
Московский государственный университет
им. М. В. Ломоносова

Киселева М. Ф. (mar-fed@yandex.ru),
Носырев Г. В. (grigorij-nosyrev@yandex.ru),
Рикитянский А. М. (nordicdx@gmail.com)
Яндекс

Васильев П. К. (asthmor@gmail.com),
Кадыкова А. Г. (annakadykova@nm.ru),
Матиссен-Рожкова В. И. (heinin@mail.ru)
Российский государственный гуманитарный университет,
Москва

В докладе рассматриваются проблемы, связанные с составлением нового словаря-тезауруса бытовой терминологии русского языка. Привлечение нового материала ставит перед лексикографами интересные задачи, требующие новых подходов. Большое внимание уделяется работе с данными Интернета, в частности для определения частотности употребления лексем и их орфографических вариантов в текстах разных жанров и в логах пользовательских запросов. Описываются инструменты, созданные авторами словаря для получения достоверных численных данных. Важное место среди активно используемых методов анализа лексики занимают опросы информантов и эксперименты. В докладе описываются их условия и анализируются полученные результаты. Результаты работы могут оказаться полезными не только в лексикографии, но и в теоретической семантике, поскольку рассматриваемый материал, ранее не подвергавшийся системному изучению, обнаруживает нетривиальные свойства.

Ключевые слова: семантика, лексикография, предметная лексика, орфография, орфоэпия, частота, статистические методы

[1] Работа выполнена при финансовой поддержке Программы фундаментальных исследований отделения историко-филологических наук РАН «Язык и литература в контексте культурной динамики», гранта РГНФ №10-04-00273а и гранта НШ-6577.2012.6 для поддержки научных исследований, проводимых ведущими научными школами РФ.

В работах [Иомдин 2009, 2011, Iomdin et. al. 2011] были рассмотрены проблемы, связанные с описанием бытовой предметной лексики, и предложен проект создания нового иллюстрированного словаря-тезауруса бытовой терминологии (СБТ). В 2010 году в Институте русского языка им. В. В. Виноградова РАН был организован и начал работу научный семинар под руководством Б. Л. Иомдина[2], на котором обсуждаются материалы будущего словаря. Словник словаря постоянно пополняется, на текущий момент он составляет более 1600 слов. Предполагаемая структура словарной статьи опубликована в [Иомдин 2011]. Настоящий доклад построен следующим образом: каждая зона словарной статьи комментируется с точки зрения возникающих проблем и предлагаемых методов их решения.

1. Группа слов с близким значением (вход)

1.1. Орфография

При составлении толковых словарей проблема орфографии обычно не является ведущей, поскольку их авторы используют данные специализированных словарей; так, словарь [Апресян 2010] опирается на [Орфографический словарь 2005]. Однако в активно используемой бытовой лексике существенную часть (по нашим подсчетам, не менее четверти) составляют новые слова, еще не зафиксированные словарями. Написание этих слов часто еще не устоялось. В условиях отсутствия признанной нормы приходится ориентироваться на узус, иногда противоречивый (ср. варианты *флэшка/флешка, пихора/пехора, ролеты/роллеты, пусета/пуссета, хайратник/хаератник*). Порой возникает вопрос, предпочесть ли более частотное написание этимологически обоснованному. Так, слово *борсетка*, восходящее к итальянскому *borsetta*, гораздо чаще пишут как *барсетка* (ср. ситуацию со словом *колготки*, восходящим к чешскому *kalhoty*, в [Левонтина 2010: 240–243]).

В последнее время для определения частотности написания часто используют данные поисковых систем (количество найденных документов). Однако такой подход ненадежен, особенно для слов, частоты которых не отличаются на порядки. Во-первых, поисковые системы расширяют запрос не только словоформами, но и синонимами, ассоциациями, транслитерацией, переводом и т. п. (что лишь частично решается использованием языка запросов). Во-вторых, количество найденных документов зависит от многих технических факторов, например, размера поисковой базы, загруженности серверов в момент запроса и т. д. Результаты по одному и тому же запросу за один период времени

[2] Кроме авторов доклада, в число основных участников семинара входят: Ф. Г. Винокуров, А. Н. Выборнова, С. А. Колмановская, А. С. Панина, А. Д. Стрижевская. Авторы хотели бы выразить признательность всем названным коллегам за активное участие в работе семинара и в подготовке материалов к докладу.

могут существенно различаться (так, запрос *портфель*, отправленный в Яндекс в декабре 2011 года с интервалом в 5 дней, выдал 428 000 и 1 000 000, соответственно). Ср. обсуждение сходных проблем в [Kilgariff 2007; Беликов, Ахметова 2009; Беликов 2011].

Помимо числа найденных документов, о частоте слова можно судить по другим источникам, например, по текстам блогов или по логам запросов пользователя (см. ниже).

Правильное написание часто определяется по поведению поисковой системы, которая может по-разному реагировать на запросы. Например, запрос *кожаная ссумка* будет исправлен на *кожаная сумка* автоматически (автозамена), на запрос *легинсы* система отреагирует подсказкой «быть может, вы искали: *ле**гг**инсы*» (подсказка), а запрос *повербол* признает безошибочным. В случае автозамены вероятность того, что введенный запрос содержит опечатку, больше, чем в случае подсказки. Если исправления не произошло, то, скорее всего, это означает, что единый вариант написания не устоялся. Чтобы увеличить надежность определения опечатки, используется не один однословный запрос, а коллекция неоднословных запросов (*кожаные ссумки*, *женские ссумки*, *дизайнерские ссумки*). Список возможных вариантов написания слова можно получить из пользовательских запросов, выбрав ближайшие по взвешенному расстоянию Левенштейна [Левенштейн 1965]. Этот метод хорошо работает на словах, в которых неоднозначны гласные (*поуэрбол*, *пауэрбол*, *повербол*).

1.2. Орфоэпия

Сходные проблемы возникают и с указанием произношения. Отметим две из них: определение места ударения (*пиала*, *боло, нострил(л)а*, *зиппо*, *унты*, *шарфы*) и произношение твердых/мягких согласных перед звуком, обозначаемым буквой *е* (*постер, дерби, трекки, лабрет(т)а*).

При определении места ударения возникают проблемы трех родов.

1) В недавно заимствованных словах в противоречие вступают две тенденции. С одной стороны, русский язык имеет склонность оформлять заимствования «колонным» ударением на последнем слоге основы, если она оканчивается на согласный, и на предпоследнем слоге, если она оканчивается на гласный (более подробно см. в [Суперанская 1968]). С другой стороны, источником большинства заимствований является английский язык, и многие носители русского языка, используя модное новое слово, произносят его с английским ударением, которое обычно падает на первый слог. Естественно, что в случае неодносложных основ часто возникает конфликт. Примером такого конфликта является слово *ноутбук*, и даже словари здесь расходятся: [Русское словесное ударение 2001] предлагает вариант *ноутбу́к*, а [Орфографический словарь 2007] — *но́утбук*. Таким образом, в этом и других подобных случаях идет борьба двух вариантов тривиального ударения (ударения, которое всегда падает на основу, причем на один и тот же по счету слог слева, см. [Зализняк 1985: 17]), падающих на разный слог основы.

2) С редко используемыми словами возникает иная проблема: конкурируют нетривиальное нормативное и тривиальное ненормативное ударение. Например, обувь *котыи́* многие склонны называть *ко́ты*; для слова *доска* в молодежной речи в последнее время отмечается ударение на первом слоге.

3) В свою очередь, хорошо освоенные слова имеют тенденцию переходить от тривиального ударения к нетривиальному (см. [Зализняк 1985: 23]), и не всегда понятно, насколько далеко этот процесс зашел для каждого конкретного слова. Неясно, например, можно ли уже признать нормативным произношение формы мн. ч. от слова *шарф* с конечным ударением (*шарфы́*) или надо по-прежнему указывать вариант *ша́рфы*.

Что же касается твердости/мягкости согласных перед *е*, она, как и ударение, коррелирует с освоенностью слова. Орфоэпические данные такого рода могут быть получены только в результате опросов: нормативные словари, как и в случае с орфографией, часто не успевают отразить новую лексику.

1.3. Словообразовательные варианты

Интересную проблему представляют существительные с диминутивными суффиксами. Как известно, уменьшительность часто лексикализуется, что требует включения таких слов в качестве отдельных входов: ср. *театральная сумочка <*сумка>*, *шапочка <*шапка> для душа*, *ремешок <?ремень> для часов*, *половая тряпка <*тряпочка>* vs. *тряпочка <*тряпка> для очков*. В процессе работы над словарем обнаружилось системное различие, особенно характерное для названий предметов одежды: уменьшительное существительное часто применимо только к детским или женским вещам (ср. *трусики — трусы*, *поясок — пояс*, *кофточка — кофта*). Кроме того, диминутивы часто характерны для женской речи (так, по нашим данным, мужчины реже говорят *колечко*, чем *кольцо*). В некоторых случаях неуменьшительное слово используется в официальной номенклатуре, ср. *брошка — брошь*[3], *батарейка — батарея* и т. п.

2. Толкование совпадающей части значения

В лексикографии традиционно признается, что исчерпывающее толкование предметной лексики неосуществимо; ср., например, [Апресян 2010: 89]. Цель нашего словаря — максимально полно представить сведения о значении и употреблении слов с предметным значением в разных стилистических пластах современного русского языка. Одна из трудностей описания значения предметных слов — указание размера. В соответствии с принципами Московской семантической школы мы стремимся к использованию природных “эталонов”. Так, размер *флешки* указывается как ‘не больше пальца’, размер

[3] Наблюдение И. Б. Левонтиной.

карточки — ‘примерно с ладонь’, размер *бумажника* — ‘примерно с кисть руки’. Но иногда найти такой эталон не удается, и более целесообразным кажется использовать артефакты: лист бумаги формата А3 для *дипломата*, паспорт для *ксивника*.

3. Близкие слова и группы слов

В СБТ принят принцип описания слов по группам, подобно синонимическим рядам в [НОСС 2004]. Деление “непрерывного пространства языка”, в любом случае искусственное, но необходимое в лексикографической практике, иногда особенно затруднительно. Одна из типичных ситуаций — наличие в значении слова компонента, конфликтующего с толкованием совпадающей части значения. Так, *чулки* — не **колготки** (являются парным предметом и не покрывают нижнюю часть туловища) и не **носки** (охватывают ногу выше колена, в основном эластичные и тонкие), но против выделения их в отдельную группу говорит тот факт, что и *носки*, и *колготки* могут служить родовым словом для *чулок*. Похожая проблема возникает с составом группы **перчатки**: стоит ли включать туда слово *муфта*? Аргументом “за” является сходство функций и материала; против — признаки ‘парность’ и ‘количество отделений для пальцев’. Таким образом, можно счесть *чулки* и *муфту* маргинальными предметами, которые, примыкая к некоторой группе, отличаются от ее доминанты сильнее других слов группы, а можно описывать эти слова отдельно, тем самым умножая количество групп.

4. Доминанта группы

В [Иомдин 2010] рассматривались различные критерии выделения доминанты в группе слов со сходным значением. Один из способов определить доминанту — эксперименты на представительной выборке носителей. Мы проводим эксперименты нескольких типов. Наиболее простой — прямое обращение к носителям с просьбой назвать предметы, изображенные на картинке (так изучались классы СУМКИ и ГОЛОВНЫЕ УБОРЫ). Более показательно описание картинки, на которой изучаемый предмет находится не в фокусе внимания испытуемого, и выявление таким образом слова, наиболее часто выбираемого для описания сходных предметов (так проводился эксперимент для групп **свитер** и **сандалии**). Еще один письменный эксперимент устроен как модифицированная игра в “банальности”. Испытуемым демонстрируется предмет и дается задание назвать его тем словом, которым его назовет большинство носителей (стимулом является получение числа очков, соответствующее числу игроков, выбравших то же слово). Так изучались слова в классах ОБУВЬ и ГОЛОВНЫЕ УБОРЫ. Это позволяет выявить самое “банальное”, по мнению большинства, название предмета, которое предположительно и является доминантой группы.

При выборе доминанты часто возникает проблема противоречия критериев ее определения, даже если ориентироваться на частотность. Так, в нормативных (и любых сколько-нибудь официальных) документах встречается практически исключительно *бюстгальтер*, а в разговорной речи безусловно лидирует *лифчик*; *сорочка* из ГОСТов и ценников соответствует повседневной *рубашке*.

Еще одна проблема — существование доминант "разных уровней".

Так, класс ОБУВЬ естественно делить на следующие группы: **сапоги**, **ботинки**, **туфли**, **кроссовки**, **тапочки**, **галоши**, **сандалии**. Однако при таком разделении возникает несколько трудностей, связанных с распределением слов по этим группам.

Внутри группы слов, объединенных одной доминантой, могут встречаться собственные «субдоминанты». Так обстоят дела с группой **сандалии**, куда входят лексемы *сандалии, босоножки, шлепанцы, шлепки, вьетнамки, сланцы*. Носители разделяют *сандалии* и *босоножки* с одной стороны, и *шлепанцы*, *вьетнамки* и *сланцы* — с другой. При этом в первом случае «субдоминантой» будет слово *сандалии*, а во втором — *шлепанцы*. Результаты опроса по картинкам, на которых летняя обувь была не в фокусе внимания, показывают, что обувь с открытой пяткой и без каблука называют *шлепанцами* приблизительно в 40% случаев (остальные употребления значительно менее регулярны), а обувь с закрытой пяткой называют сандалиями примерно в 60% случаев.

Слово *тапочки* в первом значении описывает домашнюю обувь, но у него выделяется и другое значение: 'легкая обувь на низком каблуке, предназначенная для ношения в теплое время года'. Получается, *тапочки* в этом случае могут служить гиперонимом для таких слов, как *мокасины*, *топ-сайдеры* (**ботинки**), *тенниски* (**кроссовки**), *балетки, балеринки, эспадрильи* (**туфли**), *босоножки*, *шлепанцы* (**сандалии**). Значит, это — своеобразная «супердоминанта», способная заменять слова из разных групп. Ср.: (1) *В этот самый момент в кабинет вошла женщина в форменной куртке, в фуражке, в черной юбке и в тапочках* (М. Булгаков, Мастер и Маргарита); (2) *Я вот на днях съездил в Баку, закупил партию летней обуви, шлепанцев, тапочек всяких* («Бизнес-журнал», 2004.08.17); (3) *Весна, а дворник наш уже в тапочках, в шлепках* («Русская Жизнь», 2008).

В некотором роде «супердоминантой» оказывается и слово *ботинки*, поскольку носители часто называют так мужские туфли. Ср.: (4) *Перед концертом я успел найти в магазине пару ботинок под костюм для первого отделения, а на второе — обуви не нашлось* («Вечерняя Москва», 2002.01.10); (5) *Погуляв некоторое время на улице, я снова пришла, позвонила в дверь, и мне открыл Андрей, одетый в черный костюм с бабочкой и шикарные ботинки* («Комсомольская правда», 2007.08.16). Слово *chaussure* в названии французского фильма «Le grand blond avec une chaussure noire» было переведено как *ботинок* («Высокий блондин в черном ботинке»), хотя на плакате к фильму изображена мужская туфля. Однако объединять **ботинки** и **туфли** в одну группу кажется неправильным, потому что в таком случае в группу **ботинки** войдут, например, *лодочки*, *шпильки*, *сабо*, *танкетки*, а наши опросы показывают, что носители обычно не готовы называть *ботинками* вышеперечисленные предметы обуви.

Определенную сложность представляет и описание класса ГОЛОВНЫЕ УБОРЫ. Прежде всего, само название класса принадлежит к официальному или канцелярскому стилю и редко используется в разговорной речи; вместо него обычно употребляется слово *шапки*. С другой стороны, этот класс включает в себя несколько групп, в том числе **шапки** (например, *ушанка, петушок, кепка*) и **шляпы** (например, *канотье, ток, стетсон*). Таким образом, слово *шапка* может выступать и в роли «супердоминанты», и в роли доминанты своей группы, противопоставленной *шляпе* как доминанте другой группы.

Особый случай — ситуация, когда предполагаемая доминанта является наиболее частотной и наиболее нейтральной в группе слов, называющих предметы со схожей функцией, однако ни одно из слов группы не допускает замены на другое: ср. *зажигалка, огниво, прикуриватель, спички.*

5. Различительные признаки, релевантные для данной группы

Серьезные трудности, возникающие при попытке истолковать слова с предметным значением, подробно обсуждались в [Иомдин 2009, 2011]. Привлечение нового материала выявляет дальнейшие проблемы. Так, при описании слова *сумочка* обнаружилось, что необходимо так или иначе отразить в толковании следующие смыслы: ‘небольшой размер’ (vs. *баул*), ‘элегантность, претензия на стиль’ (vs. *хозяйственная сумка*), ‘использование для транспортировки предметов, которые используются в течение дня’ (vs. *чемодан*), ‘прочный материал’ (vs. *пакет*), ‘пол обладателя’ (vs. *борсетка*). Однако при этом наличие всех названных признаков одновременно не является необходимым: *сумочкой* могут назвать и стильную дамскую сумку с украшениями, куда с легкостью помещается ноутбук (отсутствует признак ‘небольшой размер’), и небольшую дорожную сумку на пояс (отсутствует признак ‘элегантность’).

В случае, если признак действительно релевантен, далеко не всегда удается подобрать точное и общепонятное название для этого признака. Так, термин *тулья* удобен для описания шляп, но почти не известен носителям.

6. Смысловые различия

Основной способ уточнения значения названий предметов быта — опросы информантов, призванные выявить различия между названиями сходных предметов, существующие в их идиолектах.

В одном из наших экспериментов испытуемым были предложены следующие группы слов: *кувшин-графин*; *чашка-кружка*; *колье-ожерелье-бусы*; *джемпер-пуловер*; *варежки-рукавицы*; *шарф-кашне*; *кепка-бейсболка-кепи*; *платок-косынка*; *портфель-рюкзак*; *дипломат-кейс*; *маркер-фломастер*; *шпилька-невидимка*; *кошелек-бумажник-портмоне*; *банкнота-купюра*. Для каждой группы

слов хотя бы несколько испытуемых констатировали различия. Сопоставив все ответы, мы пришли к следующим выводам:

1. Существуют пары слов, для которых не удалось выявить никаких объективных различий: *джемпер-пуловер*; *банкнота-купюра; шарф-кашне*; *дипломат-кейс*. Больше половины опрошенных сходятся на том, что значения этих слов не различаются, остальные приводят различия, не сводимые к одному основанию.

2. В других случаях различия достаточно хорошо осознаются опрошенными (удалось выявить несколько регулярных различительных признаков). Это пары слов: *варежки-рукавицы; портфель-рюкзак; чашка-кружка; маркер-фломастер*.

3. Для некоторых слов четко выделяется один различительный признак, а другие многочисленные различия и близко не набирают большинства среди ответов: *кувшин-графин* ('материал'); *кошелек-бумажник-портмоне* ('назначение').

4. Для слов, входящих в группы *колье-ожерелье-бусы*; *кепка-бейсболка-кепи*; *платок-косынка*; *шпилька-невидимка*, не удается выявить «лидера» среди различительных признаков. Большинство испытуемых считают, что различия внутри этих групп есть, и описывают их. Однако ни одно из предложенных различий не набирает хотя бы половины голосов. Для одних и тех же слов внутри группы разные опрошенные приводят противоположные признаки, например, *платок-косынка*: 'квадратная форма'; *колье-ожерелье*: 'крепится под горлом, плотно охватывает основание шеи'; *шпилька-невидимка*: 'наличие декоративного элемента'. Для слов *кепка-бейсболка-кепи* различительные признаки многочисленны ('форма', 'козырек', 'назначение', 'стиль', 'пол обладателя', 'цвет', 'материал'), но ни один из них не выделяется.

Стоит также отметить следующее. В группах *кепка-бейсболка-кепи*, *кошелек-бумажник-портмоне*, *шарф-кашне* слова *кепи*, *портмоне*, *кашне* в описании чаще всего характеризовались эпитетами «изящный, изысканный, для солидных дам» (*кашне*), «гламурное, престижное» (*портмоне*), «что-то английское, экзотическое, приличное» (*кепи*) — в анкетах разных опрошенных. По-видимому, это связано с тем, что они еще не до конца освоены (морфологически это выражено в том, что слова *кепи*, *портмоне* и *кашне* не склоняются, в отличие от хорошо освоенных заимствований *кепка*, *бейсболка*, *шарф*), поэтому предмет, обозначаемый ими, приобретает коннотацию "элитарности" по сравнению с другими словами и предметами из тех же групп.

Результаты эксперимента выявляют интересную проблему: правомерно ли постулировать различия в значениях обычных слов, называющих предметы быта (не терминов!), если они отражены в нормативных источниках (толковых словарях), но не признаются или не ощущаются большинством образованных носителей? Как описывать предметы, различительные признаки которых неочевидны? Задача нашего словаря — по возможности представить весь спектр мнений носителей (соотнесенных с данными об их социальном статусе), чтобы пользователь словаря мог сопоставить эти сведения с данными нормативных

источников, приводимых в соответствующих зонах, и принимать решение в зависимости от своей коммуникативной задачи.

7. Другие значения слов, входящих в группу

Особую сложность при лексикографическом описании предметов быта представляют случаи полисемии. Так, например, лексемы слова *шлепанцы*, по-видимому, входят в группы <1> домашняя обувь с открытой пяткой (**тапочки**), <2> уличная обувь, которую носят в теплую погоду (**сандалии**), обозначая при этом один и тот же предмет. Ср.: (6) *Точно так же тело змеи трется о сухие листья, каблучки отбивают дробь по асфальту, а тапочки старого человека, которому уже трудно поднимать ноги, при каждом его шаге шлепают по полу (неслучайно одна из разновидностей мягких тапочек так и называется — шлепанцы)* (И. Иткин, «Наука и жизнь», 2006); (7) *А потом вдруг обнаружила, что они уже все втроем едут в такси, причем на ногах у Фаины были домашние шлепанцы* (Т. Тронина); (8) *Парень сплюнул, бросил сыпанувший искрами окурок и, пошаркивая пляжными шлепанцами, двинулся обратно на станцию* (А. Иличевский). Получается, что один и тот же предмет, называемый одним и тем же словом, тем не менее попадает в две группы, отличающиеся по функции.

Еще одна трудность — соотнесение в этой зоне лексем разных уровней, как, например, в случае со словом *тапочки* (‘домашняя обувь’ или ‘легкая обувь на низком каблуке, предназначенная для ношения в теплое время года’): здесь приходится описывать “супердоминанту” (см. выше) как другое значение слова, что не позволяет различие “уровней” описания.

8. Другие редкие слова, примыкающие к группе

Среди слов с близким значением, называющих бытовые предметы, неизбежно выделяются ядерная и периферийная части. Основным критерием для отнесения слов в ту или иную часть является их относительная частотность. Информацию о частотах можно получать из разных источников: частотный словарь (например, [Ляшевская, Шаров 2009]), НКРЯ, опросы информантов. Однако все эти способы обладают существенными недостатками: быстрое устаревание данных, недостаточная представленность разговорной, сленговой и т. п. лексики и недостаточность выборки для статистически значимых выводов.

Для подсчета различных словарных статистик можно использовать готовые сервисы, например: Яндекс-статистика (http://wordstat.yandex.ru/), язык запросов (http://help.yandex.ru/search/?id=1111313), Google Ngram Viewer (http://books.google.com/ngrams), Google Trends (http://www.google.com/trends/). Существуют и иные источники информации о частотности, например, поиск по блогам или текстам запросов пользователей, главные преимущества которых — стабильность и воспроизводимость результатов, замкнутость

системы (частоты слов сравнимы), большой объем и актуальность данных, а недостатки — омонимия и высокая специфичность материала.

8.1. Блоги

Поиск и подсчет статистики по Яндекс.Блогам связан с техническими трудностями, общими для любых поисков в Интернете. Отметим еще, что для получения более точного результата необходимо отделять текст поста от комментариев, что, как правило, сложно реализовать.

Разработанный внутри Яндекса специально для нашего словаря инструмент позволяет обойти все описанные проблемы и получить доступ к исходным текстам Яндекс.Блогов. Основной функцией этого инструмента является подсчет точного числа вхождений заданного слова в текстах постов с учетом всех словоформ. Итоговая частота слова складывается из частот всех его словоформ.

8.2. Логи

Пользовательские запросы (логи) являются актуальным источником разносторонней информации для лексикографа. В нашем эксперименте для подсчета частот использовались логи за месяц (декабрь 2011), обрезанные по частоте 10, что составляет порядка 12 миллионов различных запросов.

Словоформы могут иметь разную частоту, и словарная форма не всегда преобладает, поэтому под частотой слова мы понимаем сумму частот всех его словоформ[4]. Частоту слова мы ассоциировали с количеством запросов, в которых оно встретилось.

Основная проблема при определении частоты слова — развитая омонимия (так, по словоформе *гвоздики* могут быть найдены слова *гвоздик* ‘небольшой гвоздь’, *гвоздики* ‘сережки’, *гвоздика* ‘специя’, *гвоздика* ‘цветы’, *Гвоздиков* ‘фамилия’ и т. д.), снятие которой весьма трудоемко. Методы снятия омонимии, рассмотренные, например, в работах [Сокирко, Толдова 2005], [Зеленков и др. 2005], используют большие размеченные корпусы текстов, что неприменимо к нашему материалу: создание таких корпусов по трудоемкости превысило бы ручное снятие омонимии[5].

Рассмотрим предлагаемый нами алгоритм на примере двух слов: *зажим* (для денег) и *ключи* (от дома) (уточнение в скобках описывает целевой смысл). Частота каждого из этих слов порядка 10^5 и $8{\cdot}10^6$ соответственно (за один месяц). Для точного подсчета требуется разделить все пользовательские запросы,

---

[4] Для построения парадигмы слова использовалась технология mystem (http://company.yandex.ru/technologies/mystem).

[5] Отметим также, что в ряде случаев нам необходимо различать и полисемию, ср. *шлепанцы*, *тапочки*, *бахилы*.

содержащие эти слова, на три группы в соответствии с тем, в каком смысле в них употреблено слово (в целевом смысле, не в целевом смысле или затруднительно ответить). Запросы третьей группы удаляются из выборки. Поскольку размечать все запросы часто чересчур долго (скажем, для слова *ключи* их порядка 10^4), можно выбрать заданный процент либо фиксированное число самых частых запросов.

Таблица 1. Пример снятия омонимии

зажим для денег	6611	+	*скачать ключи*	1678566	–
зажим для галстука	4428	–	*горячие ключи*	60068	–
зажим купить	3765	?	*ключ смотреть*	59478	–
зажим анкерный	1947	–	*рабочие ключи*	37938	–
			ключи от квартиры	3175	+
всего запросов со словом *зажим*	94563		всего запросов со словом *ключи*	8668411	
доля целевого смысла ‘для денег’	0,4		доля целевого смысла ‘от дома’	0,002	
скорректированная частота	37058		скорректированная частота	17659	

Такой метод позволяет определить частоту слова в любом заданном значении; можно также повышать точность для отдельных слов, увеличивая количество размеченных запросов. К недостаткам следует отнести необходимость использовать ручной труд, затруднения при разметке запросов (что делать с названиями, например, *фильм «Шпильки»*?). В дальнейшем можно автоматически разметить часть запросов, используя стоп-слова (например, если в запросе со словом *зажим* есть слово *деньги*, то это скорее всего целевой смысл; если в запросе со словом *ключи* есть слово *скачать*, то это скорее всего нецелевой смысл).

Важно отметить, что тексты блогов, логов и иных документов в Интернете могут существенно различаться по сравнительной частотности употребления искомых слов. Поэтому с уверенностью ответить на вопрос, относится ли рассматриваемое слово к ядру или периферии, можно лишь в том случае, если соотношение по всем типам текстов сравнимо; в противном случае можно лишь говорить о стилистических различиях и снабжать слова соответствующими пометами.

8.3. Результаты

По полученным наблюдениям за период 01.12.2011–31.12.2011 видно, что частоты, посчитанные с помощью поисковых машин, нестабильны и часто дают неверное представление о распределении слов в Интернете (см. таблицу 2: бóльшее значение выделено полужирным шрифтом).

Таблица 2. Статистическая информация по разным интернет-источникам

Слово	Инструмент поиска по Яндекс.Блогам	Открытый поиск на blogs.yandex.ru	Инструмент по логам	Яндекс**	Google**
сапоги	**294 710**	**61 502** 61 250*	**569 271**	**571 000** **2 000 000***	2 230 000 2 180 000*
ботинки	180 019	41 089 41 079*	361 612	403 000 1 000 000*	**2 410 000** **2 380 000***
кружка	**510 153**	58 343 58 287*	**113 792**	**767 000** **2 000 000***	946 000 913 000*
чашка	422 350	**69 090** **69 068***	54 628	653 000 1 000 000*	**1 890 000** **1 840 000***

* — результат по запросу через 5 дней.
** — количество найденных документов

В зоне редких слов также может оказаться целесообразным указывать оценочные слова, чья семантика существенно размыта по сравнению с названиями конкретных предметов одежды, ср. *обдергайка*, *нахлобучка*, *кацавейка, прощай молодость* и т. п.

9. Региональные варианты

Проблемы, связанные с региональными вариантами русского языка, подробно изучаются в рамках проекта [Словарь “Языки городов”], где существенная часть словника относится к области предметной лексики; ясно, однако, что здесь предстоит еще большая работа. В проекте СБТ также планируются исследования региональных вариантов лексики с использованием данных поисковых систем и проведение соответствующих опросов информантов.

10. Данные нормативных документов

Не меньшую проблему вызывает поиск информации в нормативных документах типа ГОСТов. Это обусловлено, в частности, тем, что в таких документах принципиально иной принцип классификации предметов. Так, *галоши*, которые в словаре бытовой терминологии попадают в категорию ОБУВЬ, в ГОСТе следует искать в разделе “резиновые изделия”. Кроме того, для многих описываемых предметов нормативных документов вовсе не существует (например, для маркеров, кроссовок, перстней и пр).

11. Этимология и время появления слова

Здесь проблематична прежде всего вторая часть — время появления слова. Данных существующих корпусов недостаточно для достоверного определения этого параметра с точностью до года — можно получить лишь верхнюю оценку времени появления («Слово вошло в язык не позже … года»). И если для слов, вошедших в язык, например, в XIX веке, мы можем довольствоваться точностью в десятилетие, то для новейших слов, которые стали употребительны уже в XXI веке, такой оценки на нынешней временной дистанции недостаточно. Хорошим инструментом для оценки времени проникновения слова в книги является сервис Google Books, но большинство слов, входящих в словник СБТ, активнее употребляются в разговорной речи, чем в книжной, и поэтому важно время их появления на форумах, в блогах и т. п. Проблемы, связанные с текстами из этих источников, были рассмотрены выше.

Литература

1. *Апресян* и др. 2010 — Апресян В. Ю., Апресян Ю. Д., Бабаева Е. Э., Богуславская О. Ю., Галактионова И. В., Гловинская М. Я., Иомдин Б. Л., Крылова Т. В., Левонтина И. Б., Птенцова А. В., Санников А. В., Урысон Е. В. Проспект активного словаря русского языка. Отв. ред. акад. Ю. Д. Апресян. М.: «Языки славянских культур», 2010. 784 с.
2. *Беликов, Ахметова* 2009 — Беликов В. И., Ахметова М. В. Статистическая оценка функциональных свойств лексики по материалам интернета // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). М.: РГГУ, 2009, с. 25–130.
3. *Беликов* 2011 — Беликов В. И. Чего не хватает в «оцифрованном мире» лексикографу и социолингвисту // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2011» (Бекасово, 25–29 мая 2011 г.). Вып. 10 (17). М.: РГГУ, 2011.
4. *Зализняк* 1985 — Зализняк А. А. От праславянской акцентуации к русской. М.: Наука, 1985.
5. *Зеленков* и др. 2005 — Зеленков Ю. Г., Сегалович И. В., Титов В. А. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2005» (Звенигород). Вып. 4 (10). М.: Наука, 2005.
6. *Иомдин* 2009 — Иомдин Б. Л. Терминология быта. Поиски нормы // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). М.: РГГУ, 2009, с. 127–135.

7. *Иомдин* 2011 — Иомдин Б. Л. Материалы к словарю-тезаурусу бытовой терминологии. СВИТЕР: образец словарной статьи // Слово и язык. Сборник статей к восьмидесятилетию академика Ю. Д. Апресяна. Отв. ред. И. М. Богуславский, Л. Л. Иомдин, Л. П. Крысин. М.: «Языки славянских культур», 2011, с. 392–406.
8. *Левенштейн* 1965 — Левенштейн В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии Наук СССР, 163, №4, 1965: 845–848.
9. *Левонтина* 2010 — Левонтина И. Б. Русский со словарем. М.: Азбуковник, 2010.
10. *Ляшевская, Шаров* 2009 — Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.
11. *НОСС* 2004 — Новый объяснительный словарь синонимов русского языка. Изд. 2-е. Под общ. рук. Ю. Д. Апресяна. Москва; Вена: Языки славянской культуры: Венский славистический альманах, 2004.
12. *Орфографический* словарь 2005 — Русский орфографический словарь. Изд. 2-е, исправленное и дополненное. Отв. ред. докт. филол. наук В. В. Лопатин. М., 2005.
13. *Орфографический* словарь 2007 — Орфографический словарь. Под ред. В. В. Лопатина. М.: РАН, 2007.
14. *Русское* словесное ударение 2001 — Русское словесное ударение. Словарь нарицательных имен. Сост. М. В. Зарва. М.: ЭНАС, 2001.
15. *Словарь* «Языки городов». ABBYY Lingvo Клуб, http://www.lingvo.ru/goroda/articles.asp
16. *Сокирко, Толдова* 2005 — Сокирко А. В., Толдова С. Ю. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп) // Интернет-математика 2005. Автоматическая обработка веб-данных [Текст]: материалы о Школе. Ред. И. Сегалович, М. Маслов, Ю. Зеленков. М.: Яндекс, 2005.
17. *Суперанская* 1968 — Суперанская А. В. Ударение в заимствованных словах в современном русском языке. М.: Наука, 1968.
18. *Iomdin* et al. 2011 — Iomdin B., Piperski A., Russo M, Somin A. How different languages categorize everyday items. In: Computational linguistics and intellectual technologies. Papers from the annual international conference “Dialogue” (2011). Moscow: RGGU, 2011, p. 258–268.
19. *Kilgariff* 2007 — Kilgariff A. Googleology is Bad Science. In: Computational Linguistics, MIT Press, 33 (1), pp. 147–151.

VERBS DESCRIBING SOUNDS OF INANIMATE OBJECTS: TOWARDS A TYPOLOGY

Kashkin E. V. (egorkashkin@rambler.ru)
Moscow State University

Reznikova T. I. (tanja.reznikova@gmail.com)
All-Russian Institute for Scientific and Technical Information, Russian Academy of Sciences

Pavlova E. K. (eliz-pavlova@yandex.ru)
Moscow State University

Luchina E. S. (luchina-lena@yandex.ru)
Moscow State University

The article deals, in a typological perspective, with verbs describing sounds of inanimate objects (cf. the noise of a door being opened, of coins in somebody's pocket, of a river, etc.). The analysis is based on the data from four languages (Russian, German, Komi-Zyrjan, Khanty), which were obtained from dictionaries, text corpora and field investigation. We discuss the primary meanings of these verbs and identify the parameters that underlie semantic distinctions between them (type of sound source and its features, type of situation causing the emission of a sound, acoustic properties of sounds). Then we concentrate on the semantic shifts undergone by sound verbs. First, we consider their metonymic changes, focusing on morphological and syntactic processes accompanying these shifts. Second, we analyze metaphoric uses of sound verbs, bringing out typical patterns of their derivation. These results should form the basis for a future large-scale typological investigation of sound verbs.

Key words: lexical typology, semantics, sound verbs, semantic change, metaphor, metonymy

1. Introduction

It is common knowledge in modern linguistics that language facts are better described and explained not in an isolated way, but in a broader typological perspective. This holds also for lexical semantics. When working inside the system of a sole language, one may provide detailed semantic descriptions of every single lexeme, compare items close in meaning, bring out differentiating parameters. The most felicitous example of this strategy is represented by lexicographic projects lead by Ju. D. Apresjan (see NOSS, PAS). Revealing quite subtle semantic features of separate items, this approach is however not aimed at describing the system as a whole — indeed, one can understand the specificity of the way a particular semantic zone is organized in only by comparing it with an analogous zone in other languages. It is such a comparison that shows which

features of a system reflect universal cognitive principles, what corresponds to typologically frequent patterns and what to rare ones. Finally, a challenging task is to reveal what systems never occur in languages and what motivates the restrictions.

All these issues are addressed by lexical typology (cf. Rakhilina, Plungian 2007, Koptjevskaja-Tamm 2008, Rakhilina, Reznikova 2011). Within the cross-linguistic approach (like within the intralingual analysis in the tradition of the Moscow semantic school) the object of direct investigation is constituted by separate semantic fields, as every field corresponds to its own fragment of the extralinguistic world and is therefore characterized by a particular set of relevant parameters. At the same time it seems probable that some of these parameters will be found common for a substantial number of semantic fields, it may even turn out that some features will prove universal, being in effect in all the lexical zones. Such generalizations will however be reasonable only when many dozens of heterogeneous semantic domains will be thoroughly described. Nowadays this work is in fact at the very outset — large-scale typological investigations have covered only a few zones for the time being (cf. Viberg 1984 on perception verbs, Newman (ed.) 1998 on verbs of giving, Newman (ed.) 2002 on verbs of sitting, standing, and lying, Maisak, Rakhilina (eds.) 2007 on aqua-motion verbs, Britsyn et al. (eds.) 2009 on pain predicates, Krugljakova 2010 on rotation verbs, Koptjevskaja-Tamm (ed.) forthc. on temperature terms, and several others).

A semantic domain is obviously the more interesting for typology the more opportunities it provides for lexical oppositions. One of such "rich" fields is the zone of sound verbs discussed in the present article. This zone may be further divided into smaller subzones — and here the following theoretical question immediately arises: how should a lexical system be divided into fragments, i. e. which subzones should be treated within a single typological description, and which of them should be studied separately? When searching the solution of this problem, we rely not so much on ontological classification, as on language data, notably, if two semantic domains which may be in theory opposed to each other are regularly united in the same lexemes, we do not separate them, constructing a single typology of them. On the contrary, if the majority of languages lexically differentiate two zones, we treat each of them separately. In accordance with this criterion the zone of sound verbs may be divided into at least three subzones: sounds of human voice (see the article by A. V. Ptentsova in PAS (pp. 619–659) on this group in Russian), sounds of animals (see a typological overview in Rakhilina 2010), and, finally, sounds of inanimate objects. In this article we will focus on the latter subclass.

Note that the chosen subzone is quite large — thus, our verb list for Russian contains 55 lexemes — and this is not surprising, as sounding of inanimate objects correlates with a considerable number of various situations. Indeed, sounds are emitted not only by mechanisms when functioning, or by natural objects like water in a river, but virtually by any inanimate object affected by a human or by natural force (e. g., tree branches during the wind, a floor under one's feet, a door being opened, keys in the pocket of a walking man, a falling coin, heavy objects when moving, etc.). How is all this variety of sounds distributed among lexemes in different languages, which parameters of a sound situation play a crucial role in this process — these are the main questions to be addressed in typological research.

Our language sample analyzed for the moment comprises four languages: Russian (Indo-European → Slavic), German (Indo-European → Germanic), Komi-Zyrjan (Fenno-Ugric → Permic; Izhma dialect of the village of Muzhi), Khanty (Fenno-Ugric → Ob-Ugric; the dialect of the village of Tegi, intermediate between the dialects of Shuryshkary and Kazym). The sample will be enlarged at the next stages of the investigation, however the data collected so far do enable us to outline some crucial points regarding the typology of sound verbs and the perspectives for its development.

It should be pointed out that Russian sound verbs have already become the subject of several investigations (see Stojnova 2008 and references therein, as well as Paducheva 2004 and some lexicographic articles from NOSS). Our analysis of Russian data relies on both those papers and our own data acquired from the Russian National Corpus. The German data has been retrieved from dictionaries (Duden, DWDS, Wortschatz Leipzig) and text corpora (DWDS, COSMAS II), as well as obtained from native speakers. The data of Khanty and Komi-Zyrjan has been gained during their complex field investigation.

The structure of our paper is as follows. In Section 2 we discuss the primary meanings of sound verbs and the semantic parameters that underlie lexical oppositions within the field. Next sections deal with secondary meanings and the mechanisms of their development: first, we analyze metonymic (Section 3), and then metaphoric shifts (Section 4). Finally, in Section 5 we draw conclusions.

2. Sounds of inanimate objects: lexical oppositions within the field

When dividing the integral domain of sounds into three subzones, we have already hinted that the classification of sound verbs is largely dependent on the type of sound source (cf. human, animal, or inanimate). It is reasonable to assume that sound source is also involved into further lexical differentiations, i. e. within the domain of inanimate sounds.

Whereas the type of source is a common parameter for all the sound verbs, there is one more parameter specific for sounds of inanimate objects. As has already been mentioned, inanimate objects (as distinct from people and animals) typically emit sounds not on their own account, but being affected by human or natural force. It is the type of such an action (i. e. a situation leading to sound emission) that forms the second parameter lexically differentiating "inanimate" sounds.

These two parameters — the type of a source and the type of an action — on the one hand, characterize a sound situation on completely different grounds, but, on the other hand, are not fully independent, because different objects are prototypically affected in a different way. Moreover, some sound situations do not correlate with the idea of an external action at all: those are natural sounds (cf. thunder, murmur of a stream, etc.). The specific character of these sound situations also becomes apparent in their lexicalization: many natural sounds are conceptualized by separate lexemes (cf. Russian *žurčat'* and German *plätschern*, which both denote nothing but the sound of water flowing in a stream, or Komi-Zyrjan *gymoony* describing only thunder). This might give an idea that, in accordance with the criterion discussed in Part 1, natural objects should be separated from

artifacts and studied as a distinct semantic domain. In most cases, however, languages do not oppose natural sounds to those of artifacts, due to the fact that many of them turn out to be perceived as very similar. For example, the sound of dry leaves under one's feet or during a slight wind tends to be described by the same verb as the sound of papers sorted out by a person (cf. Russian *šelestet'*, German *rauschen*), and flowing water in Komi-Zyrjan, contrary to such verbs in Russian in German, 'sounds' in the same manner as an engine or a tractor (*žurgyny*). As a consequence of this, we consider both natural and artifact sounds within a single typological description (but at the same time we obviously reserve separate nodes for thunder or flowing water in our classification of sound types)[1].

Leaving apart sounds related to natural phenomena as a particular type of sound situations, let us now focus on the core part of the semantic field concerned — objects producing sounds as a result of an external action over them. We go on to consider how our parameters — the type of a sounding object and the type of an action over it — are reflected in the lexicalization of sound situations.

To begin with the type of a source, note, that the classification of sources for the other subzones of sound verbs is quite obvious. Thus, it is enough to go over various species of animals in order to check which of them are described by special sound verbs. For inanimate objects the categorization is much less obvious. It is difficult to imagine using a separate verb for every single object, so there must be some distinctive semantic features that determine "the manner" they produce sounds. According to our data, the most relevant properties of an object are the **material** and the **size / weight** of an object.

There are several kinds of material relevant for the type of sound, and their sounds are lexicalized in all the languages in our sample: metal (cf. Russian *ljazgat'* 'clank', *skrežetat'* 'grind, gnash'; Khanty *s'ărɨtɨ* 'squeak — about a rusty doorlock or an iron hinge'), glass (cf. German *klirren* or Komi-Zyrjan ideophonic verbs with the root *z'il'-*), and wood (cf. German *knarren*, Khanty *šɨxartɨ* 'creak'). Along with their prototypical objects, these verbs may cover some "external" frames, and languages show a great diversity in those peripheral zones. Thus, in Russian sounds of teeth tend to be categorized in the same way as sounds of metallic objects (cf. *ljazgat', skrežetat'*), although in some situations a more "wooden" verb *skripet'* is also possible. In Khanty, however, teeth may produce only a "wooden" sound (*šɨxartɨ*)[2].

Besides material, important features of a sound source are its size and weight: some verbs may 'specialize' either in small and light (cf. German *klimpern* describing collisions of small stones, coins, keys, etc.) or in big and heavy (cf. Russian *gromyhat'* about the strokes of heavy chains, tram, furniture) objects.

---

[1] There is still a semantic class among the sounds of inanimate objects which is lexically opposed to the other sound verbs and therefore might be the object of a separate typological investigation — we mean the sounds of musical instruments. For them, the emission of a sound is not a side effect of an action over them (cf. squeak of a door or noise of furniture being moved), but their main function. That is why languages often use special lexical units for this type of a sound. Interestingly, languages seem to most notably elaborate sounds of bad and unpleasant playing, cf. Russian *pilikat'* (about strings), *brenčat'* (about piano or guitar); German *quitschen* (about strings), *klimpern* (about piano).

[2] As regards collisions of teeth in Khanty, they are denoted by the verbs *seŋktɨ* 'to knock, to beat' and *tarɨtɨ* 'tremble', which are rather verbs of physical action or state than verbs of sound.

The size of an object may turn out to be more relevant than its material. Thus, small metallic objects (keys, coins), judging by the languages from our sample, sound rather like glass than like metal (cf. Russian *zvenet'*, Komi-Zyrjan ideophonic verbs with the root *z'il'*-, both refer to collisions of glass and small metallic objects).

The physical parameters of a sound source are further overlapped by the second parameter — the type of the situation which causes an object to produce a sound. We may mark out **falling down** and different types of **deformation** among the most typical situations leading to the emission of sound. Thus, a sound associated only with falling down is described by a German verb *plumpsen* or Komi-Zyrjan *butkys'ny* meaning 'to fall with noise'. An example in which loud deformation is lexicalized is the Komi-Zyrjan ideophonic verb *čažvartny* 'to tear a flat object with noise' (cf. the neutral *kos'oony* 'to tear a flat object'). A subtype of deformation is chewing / gnawing an object, which may also be relevant for sound verbs. For example, a Khanty verb *mŭrnăλtɨ* means 'to crunch sth. (e. g., dried crust)', cf. neutral *pŏrtɨ* 'to gnaw sth.' not implying emission of a sound. Similarly, a Komi-Zyrjan verb *rusjyny* describes a cow or a horse chewing fodder noisily (cf. neutral *n'ačkyny* 'to chew').

The sound classification according to the types of sources and causing situations allows us, on the one hand, to see what can be lexically distinguished within the class of sound verbs, and on the other, to account for what can be lexically merged into one lexical unit. Thus, in Russian both a chair and a door may *skripet'* 'creak', because they (or parts thereof) are made of wood. In German, quite different objects (cf. apples, a bag, a person) may *plumpsen* as long as the noise they produce is associated with falling. However, there are still other combinations of sound types, which are described by a single lexeme, but whose merger cannot be explained by the parameters presented so far. Cf. the above-mentioned Komi-Zyrjan verb *žurgyny* describing sounds of both a stream and a tractor; or Khanty *šŭl'itɨ*, which denotes, on the one hand, rustling of tree leaves, paper, etc., and, on the other hand, the sound accompanying collisions of small objects (coins, fragments of broken glass, bunch of keys, etc.); or again Khanty *lŏtɨtɨtɨ*, which is the basic verb for discrete crackling (e. g., that of dry wood in the fire), and at the same time describes the sound of frequent collisions (e. g., when a seagull flaps its wings). The analysis of such cases (including but not limited to those enumerated here) leads us to the idea of one more parameter that is relevant to sound conceptualization, namely its acoustic properties formulated in terms of its continuity and regularity. With respect to this, most sounds of inanimate objects may be located on a sort of **continuum**: continuous regular monotonous sounds (e. g., drone of a plane making a landing) — continuous irregular sounds (e. g., rustling of tree leaves) — continuous discrete sounds (e. g., crackle of dry wood) — regular discrete sounds (e. g., clatter of heels) — instantaneous sounds (e. g., a sound of sth. falling down). Our material demonstrates that verbs of sound in all the languages concerned cover an uninterrupted zone on this continuum (as for the examples above, those are continuous irregular sounds for Komi-Zyrjan *žurgyny*, continuous irregular & discrete irregular frequent sounds for Khanty *šŭl'itɨ*, irregular discrete & regular discrete frequent sounds for Khanty *lŏtɨtɨtɨ*). The acoustic properties by no means cancel the oppositions regarding the types of a source and of a situation in general, but they impose some additional restrictions on the 'area', in which those oppositions are applied.

3. Metonymic shifts

The extralinguistic connection between sounds of inanimate objects and situations causing their emission determines one more (purely linguistic) feature of the verbs in question: they often develop metonymic meanings, and this tendency seems to represent a linguistic universal. Indeed, if the process of emitting a sound is contiguous to a physical action (for example, a door is squeaking when it is opening, and wheels are knocking when the transport is in movement), then sound verbs may easily shift to those contiguous semantic classes, denoting a physical situation that is accompanied by a sound. This new meaning is thus metonymically related to the source one.

Such proneness to metonymy makes the domain of verbs referring to the sounds of inanimate objects a fertile ground for studying the techniques of a metonymic shift. Note that the theory of metonymy have concentrated to a greater extent on regular patterns of such shifts, but not on their linguistic mechanism (see however Paducheva 2004). Our task here is to show what linguistic means are applied by languages from our sample for expressing metonymic meanings.

The first and the more obvious strategy is the addition of an argument to a construction (see Paducheva 2004, Stojnova 2008 for details about Russian). Compare examples (1a-b)

(1) a. *Bumaga*$_{nom}$ ***šuršit*** — The paper is rustling.
b. *Mal'čik*$_{nom}$ ***šuršit*** *bumagoj*$_{instr}$ — The boy is rustling the paper.

Example (1a) contains an intransitive clause describing a sound. As regards example (1b), it gets an agentive participant (*mal'čik* 'the boy') marked with nominative, whereas the NP denoting the source of a sound changes its syntactic function and takes an instrumental affix.

Another example related to this group is (2a-b) from German:

(2) a. Als er fiel, hat es wirklich **geplumpst** — When he fell down, there was some noise.
b. Der Sack **plumpste** auf den Boden — The bag fell on the floor with noise.

The verb *plumpsen* initially denotes the sounds accompanying situations of falling (cf. 2a). In (2b) it is however used in a construction typical of a verb of falling (and, broader, of a verb of movement), which includes its combining with a locative (directional) argument *auf den Boden* 'onto the floor'.

An additional possibility to the syntactic alternations that has just been mentioned consists in the changes in morphological markers, depending on the type of construction where a verb is included. This is a frequent case for Komi-Zyrjan, where most of the verbs denoting sounds are derived from ideophones and take different affixes of aspect and actancy derivation in different constructions. For example, an ideophonic root *rač- (račk-, račča-)* refers to a sound (cracking, crunching) which accompanies breaking a wooden object or thin ice. This root forms a base for such verbs as *račk'-ed-ny* (crack-tr-inf) 'to break sth. with cracking, crunching', *račk'-ed-l-yny*

(crack-tr-iter-inf) 'to break sth. with cracking, crunching several times', *račk'-ed-č-yny* (crack-tr-detr-inf) 'to crack, crunch' (denotes a single action, which is typical of the combination of a transitive affix and an intransitive one in Komi-Zyrjan verbal morphology), *raččakyyny* 'to crack, crunch' (intransitive; *kyyny* is a grammaticalized verb 'to hear, to be audible'), *račvartny* 'to break sth. with cracking, crunching in a very intensive manner' (where *vartny* is a grammaticalized verb 'to beat' which typically conveys the idea of intensiveness in such patterns, which are not limited to verbs of sound and also include, for instance, verbs of physical action). It should be pointed out, however, that such morphological changes are not restricted to "ideophone-based" systems like that of Komi-Zyrjan. Thus, in the case of Russian *hrust'et'* 'to crunch' the metonymic shift to the domain of punctual action is marked by a semelfactive suffix *-nu-* (3a-b), cf. also a similar pair *tr'eščat'* — *tr'es**nut'*** 'to crack' where the semelfactive verb doesn't seem to denote a sound any longer and is limited to the domain of punctual deformation.

(3) a. *Vetki hrusteli pod ego nogami* — The twigs were crunching under his feet.
b. *Vetka hrust**nu**la* — The twig crunched.

Languages that use analytic forms (like German in our sample) appear to express constructional changes by choosing different auxiliaries, like in (4a-b).

(4) a. *Der Boden **hat** unter seinen Füßen geknackt* — The floor has cracked under his feet.
b. *Die Fensterscheibe **ist** geknackt* — The window glass broke (lit. 'is cracked').

In (4a) the verb *knacken* refers to a sound, and it is conjugated with the auxiliary *haben*, which is used for most German verbs. In (4b), however, the same verb stands with *sein* which is in particular taken by the verbs denoting the change of a state. This clearly shows that *knacken* in (4b) does not refer to a sound, but to a change of state, i. e. the destruction of the window.

4. Metaphoric shifts

The task of a typological research into metaphors is quite challenging, as the process of their development is creative, and consequently it may seem difficult to impose restrictions on it, and also to distinguish in all cases between a stable metaphor and an occasional one. This task is however undertaken in lexical typology (cf. Rakhilina 2007 on the verbs of aqua-motion, Britsyn et al. (eds.) 2009 on the expressions of pain, Krugljakova 2010 on the verbs of rotation, Rakhilina 2010 on the verbs of animal sounds, see also Zaliznjak 2009), because it turns out possible to identify typologically consistent metaphors and to explain the semantic motivation for their emergence, therefore enlarging the empirical basis of the theory dealing with semantic shifts.

As far as metaphors developed by verbs of sound are concerned, we propose to classify them into two basic types, dependent on the part of meaning which triggers the semantic shift. The metaphors of the first type (we call them 'acoustic metaphors')

are based on physical parameters of a sound, whereas the metaphors of the second type ('non-acoustic metaphors') come from other parameters of a situation.

Acoustic metaphors often occur in expressions dealing with the properties of smb's voice or speech. Thus, a verb *gremet'* originally means in Russian 'to thunder' or 'to clatter, clank' and develops a metaphor 'to thunder, roar' (about smb speaking loudly), coming from the idea of loudness. Another interesting pattern is a metaphorical shift of verbs describing very frequent sounds to the domain of fast speech: for example, such as the case of a Russian *treščat'* 'to crack, crackle' and of a Komi-Zyrjan *tark'edčyny* 'to knock, chatter in a frequent manner' (e. g. when one is knocking at the door, or when one's teeth are chattering because of frost) which both develop the meaning 'speak quickly'. Besides, many acoustic metaphors belong to the target domain of unpleasant physical sensations (see also the typological overview of such a shift in Britsyn et al. (eds.) 2009), cf. the Russian verb *gudet'* which denotes the monotonous sound of wind or of a plane making a landing and is metaphorically used for an unpleasant sensation in one's legs when one is tired, or for a headache. Another example of this kind is represented by the German verb *dröhnen* which primarily denotes the dinging of a bell or the noise of a window during a storm, and metaphorically describes a headache.

As regards non-acoustic metaphors, they are not motivated by a sound itself, but by a situation that leads to sound emission (cf. Part 3). Examples of this kind are widespread in Komi-Zyrjan, which develops a rich system of verbs denoting an action and at the same time the sound accompanying it. Thus, for instance, a verb *br'ingys'ny* 'to fall with noise' (about a small metallic object; cf. the neutral *us'ny* 'to fall') is used figuratively in the meaning 'to fall (about a drunk person)', therefore maintaining the non-acoustic idea of falling. Similarly, a word *č'ažvartny* 'to tear a flat object with noise' (cf. the neutral *kos'oony* 'to tear a flat object') is metaphorically applied to a person who is making abrupt movements when playing the accordion (so to speak, tearing it, which is the motivation for this metaphor). Another example in this group is Russian *treščat'* 'to crack, crackle': one of its metaphors is 'to have a headache' (the semantic extension is based here on the idea of destruction, as the sound denoted by *treščat'* typically accompanies breaking an object).

5. Conclusion

In our article we have focused on one subzone within a larger domain of sound verbs, namely on sounds of inanimate objects. Their special nature stems from a basic feature of their sources: inanimate objects do not usually produce sounds on their own account, they "sound" only as a result of an external action, i. e. of the situation they are involved in. This extralinguistic fact determines many different linguistic properties of these verbs.

Firstly, the way of describing a sound depends on what happens with the source object, i. e. the type of a situation forms one of the parameters organizing the structure of lexical oppositions within the domain. Secondly, the contiguity between sounds and physical actions favours the development of metonymic uses, and

typological data make it possible to trace the mechanism of these shifts. In particular, it is significant from the point of view of the semantic theory that a metonymic shift may be encoded not only in syntax (by a change in argument structure), but also in morphology (by means of word-building). Thirdly, the contiguity with physical actions expands the range of metaphoric shifts: metaphors may develop on the basis not only of a sound meaning proper, but also of situations causing the emission of these sounds. Thus, a detailed investigation of verbs denoting sounds of inanimate objects may contribute to the analysis of quite a few adjacent domains — verbs of cutting & breaking, falling down, etc.

Generally speaking, a typological investigation of any new semantic field should start, so to speak, from scratch — a new fragment of the extralinguistic reality involves new parameters of lexical variation. At the same time the methodology of revealing the points where lexical oppositions are potentially possible is already clear: it is necessary to understand what semantic valencies are typical of the domain concerned, and to check what types of participants filling in these valencies may be described with different lexemes. By combining different participants we set typical simple situations — frames (cf. for our domain: the sound of a floor somebody is walking on; the sound of tree leaves during the wind; the sound of a glass breaking down, etc.) Such frames are, on the one hand, easy to use as entries for a typological questionnaire, which will be applied in studying a broader language sample. On the other hand, in the future, frames could form the basis for multilingual dictionaries of a completely new type.

As is well-known for anyone who has ever used a dictionary, a word often has several translations, and it is usually difficult to understand the difference in their use from dictionary examples. If a dictionary were based on typical simple situations, and the list of these situations covered all the cases of lexical oppositions within a substantial language sample, the search of interlanguage correspondences would not take a lot of effort. This approach seems to have considerable potential of putting lexical typology into lexicographic theory and practice.

References

1. *Britsyn, Viktor M.; Rakhilina, Ekaterina V.; Reznikova, Tatiana I.; Javorska, Galina M.* (eds.) (2009), Kontsept boli v tipologicheskom osveshchenii [The concept of PAIN in a typological perspective]. Kiev, Dmitri Burago's Publishing House.
2. *COSMAS* II. Corpus Search, Management and Analysis System, availavle at: http://www.ids-mannheim.de/cosmas2/
3. *Duden* (1999), Das große Wörterbuch der deutschen Sprache in 10 Bänden. 3. Auflage. Mannheim, Dudenverlag.
4. *DWDS*: Digitales Wörterbuch der deutschen Sprache. Berlin-Brandenburgische Akademie der Wissenschaften, available at: http://www.dwds.de
5. *Koptjevskaja-Tamm M.* (2008), Approaching Lexical typology, in Vanhove, Martine (ed.) From polysemy to semantic change: Towards a typology of lexical semantic associations. Amsterdam, Philadelphia, Benjamins, pp. 3–52.

6. *Koptjevskaja-Tamm M.* (forthc.), The Linguistics of Temperature. Amsterdam, Philadelphia, Benjamins.
7. *Krugljakova V. A.* (2010), Semantika glagolov vrashchenija v tipologicheskoj perspektive [Semantics of rotation verbs in a typological perspective]. Avtoreferat dis. … kand. fil. nauk. Moscow.
8. *Maisak T. A., Rakhilina E. V.* (eds.) (2007), Glagoly dvizhenija v vode: leksicheskaja tipologija [Verbs of AQUA motion: lexical typology]. Moscow, Indrik.
9. *Newman, J.* (ed.) (1998). The Linguistics of Giving. Amsterdam & Philadelphia, John Benjamins.
10. *Newman, J.* (ed.) (2002). The Linguistics of Sitting, Standing, and Lying. Amsterdam & Philadelphia, John Benjamins.
11. *NOSS* (2004): Novyi ob"jasnitel'nyj slovar' sinonimov russkogo jazyka. Vtoroj vypusk [New Explanatory Dictionary of Russian Synonyms: Second issue], under the general supervision of Ju.D. Apresian. Moscow, Jazyki slavjanskoj kul`tury.
12. *PAS* (2010), Prospekt aktivnogo slovarja russkogo jazyka [Prospectus of an active dictionary of Russian], under the general supervision of Ju.D. Apresian. Moscow, Jazyki slavjanskoj kul`tury.
13. *Paducheva E. V.* (2004), Dinamicheskie modeli v semantike leksiki [Dynamic models in the semantics of the lexicon]. Moscow, Iazyki slavianskoj kul'tury.
14. *Rakhilina E. V.* (2010), Moo sounds [Zvuki Mu], in Problemy grammatiki i tipologii. Sbornik statej pamjati V. P. Nedjalkova [Issues in grammar and typology. A memorial volume for Vladimir Nedjalkov]. Moscow, Znak.
15. *Rakhilina E. V., Plungian V. A.* (2007), On Lexical Semantic Typology [O leksikosemanticheskoi tipologii], in Maisak T. A., Rakhilina E. V. (eds.) Glagoly dvizheniia v vode: leksicheskaia tipologiia [Verbs of AQUA motion: lexical typology]. Moscow, Indrik.
16. *Rakhilina E. V., Reznikova T. I.* (2011), On the activities of the Moscow lexical typological group [O rabote Moskovskoi leksiko-tipologicheskoi gruppy], in Problemy leksiko-semanticheskoi tipologii [Issues in lexical semantic typology]. Voronezh, Voronezh State University, pp. 126–135.
17. *Stojnova N. M.* (2008), Semantika i morfosintaksicheskie svoistva glagolov zvuka v russkom iazyke [Semantics and morphosyntactic properties of sound verbs in Russian]. Master's thesis. Moscow, Moscow State University.
18. *Viberg, Å.* (1984). The verbs of perception: A typological study. Linguistics, 21-1, pp. 123–162.
19. *Wortschatz* Leipzig, available at: http://wortschatz.informatik.uni-leipzig.de.
20. *Zaliznjak, Anna A.* On the notion of semantic shift [O poniatii semanticheskogo perekhoda]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2009"]. Bekasovo, 2009, pp. 107–112.

ОПТИМИЗАЦИЯ МОДЕЛИ РЕФЕРЕНЦИАЛЬНОГО ВЫБОРА, ОСНОВАННОЙ НА МАШИННОМ ОБУЧЕНИИ

Кибрик А. А. (aakibrik@gmail.com)
Институт языкознания РАН

Линник А. С. (skylinnik@gmail.com),
Добров Г. Б. (wslc@rambler.ru),
Худякова М. В. (mariya.kh@gmail.com)
МГУ им. М. В. Ломоносова

В статье обсуждаются различные способы оптимизации системы, моделирующей референциальный выбор (РВ) на основе аннотированного корпуса с использованием машинного обучения. Аннотационная схема, использовавшаяся в наших более ранних исследованиях, была улучшена и расширена. На следующем этапе был имплементирован более «дешевый» набор параметров с целью сокращения времени обработки и трудозатратности аннотации. Наши результаты свидетельствуют о том, что, несмотря на возможность исключения наиболее «дорогих» факторов при моделировании РВ, лучшая аккуратность предсказания достижима только при использовании максимального количества доступной информации. Жанровая принадлежность текстов была введена в систему в качестве одного из параметров и послужила повышению показателя аккуратности. И наконец, была запущена серия психолингвистических экспериментов по изучению категоричности выбора, совершаемого говорящими/пишущими. Первые полученные нами результаты оказались многообещающими: они показали, что в случаях, в которых системе не удается дать однозначное предсказание, согласно человеческой оценке, возможно с равной вероятность использование более одного референциального средства.

Ключевые слова: референциальный выбор, риторическая структура, референция, компьютерное моделирование, машинное обучение, аннотированный корпус

Введение

Целостность и состоятельность производимого дискурса (текста/речи) напрямую связана с повторным (многократным) упоминанием одних и тех же сущностей — референцией. В зависимости от того, какое референциальное средство использует говорящий — имя собственное, местоимение или дескриптивную именную группу (см. пример (1)), — может меняться как смысл высказывания в целом, так и степень доступности его смысла для адресата.

(1) *Mr. Akerson said MCI recorded "another solid cash positive quarter," its fourth in a row, but declined to comment on whether the company is considering a dividend or is planning any acquisition. "The current quarter," he said, "looks fine".*

Настоящее исследование посвящено изучению механизмов, используемых говорящим при выборе способа отсылки к референту, называемом *референциальным выбором* (РВ). Известно, что за этот процесс ответственна кратковременная память, в частности, *активация* того или иного референта в сознании говорящего (Kibrik 1996, 1999, 2011). Именно эта взаимосвязь лежит в основе главного принципа референциального выбора в модели, предложенной и разработанной А. А. Кибриком. Он заключается в том, что при низкой активации референта говорящий (или пишущий) использует наиболее лексически полные выражения, в то время как для высоко активированных референтов обычно используются более краткие формы. Данное исследование проводилось на материале английского языка (подробнее в разделе «Корпус RefRhet»), в котором основным редуцированным референциальным средством является местоимение (нули используются в крайне ограниченном количестве конструкций, а указательные местоимения чаще всего используются для ситуативной референции). Полными же референциальными средствами считаются имена собственные и определенные дескрипции.

Моделирование процесса референциального выбора на данный момент является одной из основных задач в области синтеза естественного языка. При построении модели РВ используются два варианта данной задачи: (а) выбор между редуцированным и полным референциальным средством, и более сложный трехчастный (б) выбор между местоимением, определенной дескрипцией и именем собственным. Кроме высокой точности предсказания средства отсылки к референту, необходимо наличие возможности имплементации его в общую модель порождения дискурса. В настоящем исследовании референциальный выбор моделируется при помощи машинного обучения на корпусе текстов большого объема.

Корпус RefRhet

Корпус RefRhet был создан на базе известного англоязычного корпуса RST Discourse Treebank (http://www.isi.edu/~marcu/discourse/Corpora.html), состоящего из тестов Wall Street Journal и получившего под руководством Д. Марку (Carlson et al. 2003) разметку в соответствии с Теорией Риторической структуры, далее — ТРС (Mann & Thompson,1987; Taboada & Mann, 2006). Результаты более ранних исследований (Fox, 1987; Hobbs, 1985; Kehler, 2002; Kibrik, 1996) указывают на наличие зависимости между дискурсивной структурой и референциальным выбором. Наиболее логично среди различных моделей семантико-дискурсивной структуры текста (например, Hobbs, 1985; Joshi, Prasad & Miltsakaki, 2006; Miltsakaki, Prasad, Joshi & Webber, 2004; Polanyi, 1985; Wolf & Gibson, 2003) иерархическую организацию дискурса описывает ТРС. Выбор RST Discourse Treebank обоснован наличием в нем аннотации иерархической

структуры, которая требует специальной подготовки аннотаторов и значительных затрат времени. В корпусе содержится 176 383 словоупотребления и 21 789 элементарных дискурсивных единиц (ЭДЕ). Риторическое расстояние, которое представляет собой длину пути между фрагментами текста по построенной риторической сети, считается важным фактором при референциальном выборе (Fox 1997, Kibrik 1996), поскольку позволяет учесть связь между фрагментами текста, далекими друг от друга по линейному расстоянию, но близкими по структуре изложения.

Корпус RefRhet был создан путем нанесения на RST Discourse Treebank референциальной разметки. Между всеми референциальными выражениями (markable, далее маркабула), устанавливаются отношения кореферентности — каждое непервое упоминание референта (анафор) связывается с предшествующим ему упоминанием (антецедентом). Разметка осуществлялась при помощи так называемой аннотационной схемы (подробно об исходной схеме разметки RefRhet см. Krasavina and Chiarcos 2007) — набора признаков (грамматическая роль, одушевленность др.), приписываемых каждой маркабуле, которые могут оказывать влияние на РВ, с использованием программного инструмента MMAX-2 (http://mmax2.sourceforge.net/). Каждый текст был размечен двумя независимыми аннтаторами. Мы применяли различные алгоритмы машинного обучения к данному корпусу, в результате чего была достигнута аккуратности предсказания референциальной формы до 90% для двухчастной задачи и около 80% — для трехчастной. Подробные отчеты содержатся в (Kibrik et al. 2010) и (Loukachevitch et al. 2011).

В целях оптимизации процесса аннотации было принято решение модифицировать аннотационную схему, что и было произведено к февралю 2012 года. Подкорпус RefRhet, названный RefRhet2, был аннотирован по новой схеме, которая была названа MoRA (Moscow Referential Annotation). В настоящей работе представлены результаты обработки RefRhet2 , количественные характеристики которого указаны в Таблице 1.

Табл. 1. Количественные характеристики корпуса RefRhet2

Параметр	*Количество в корпусе*
Количество слов	45 016
Количество ЭДЕ	5497
Количество маркабул (референциальных выражений)	11 461
Количество пар «анафор-антецедент»	3692
Количество референциальных цепочек (последовательностей упоминаний одного и того же референта)	1511

Методы машинного обучения

Для экспериментов были выбраны несколько алгоритмов машинного обучения, относящихся к разным типам: логические алгоритмы классификации

(деревья решений C4.5, алгоритм решающих правил JRip), результаты работы которых легко интерпретировать, а также логистическая регрессия, позволяющая получить оценки вероятности принадлежности каждому из классов и качество работы которой превосходит качество работы логических алгоритмов. Нами также были использованы композиции классификаторов: баггинг (bagging) и бустинг (boosting). Подробнее работа данных алгоритмов уже обсуждалась нами ранее (Loukachevitch et al. 2011). Критерием выбора наилучшего набора признаков и алгоритма является *аккуратность* — отношение правильно предсказанных типов референциальных выражений к их общему количеству.

Результаты обработки корпуса RefRhet2 приведены в Табл. 2.

Табл. 2. Результаты работы алгоритмов машинного обучения на корпусе RefRhet2

Алгоритм	**Аккуратность классификации для двухклассовой задачи**	**Аккуратность классификации для трехклассовой задачи**
Решающие правила	85.9%	75.7%
Деревья решений	87.1%	76.8%
Логистическая регрессия	87.9%	77.6%
Баггинг	88%	78.6%
Бустинг	88.7%	78.7%

Результаты работы некоторых алгоритмов (напр., лог. регрессии и деревьев решений) улучшились с изменением аннотационной схемы по сравнению с результатами, опубликованными в (Loukachevitch et al., 2011), однако по другим алгоритмам показатели оказались несущественно ниже, чем в прошлых исследованиях. Это связано с тем, что новая аннотационная схема включает в себя «сложные» случаи, такие как групповая аннотация и неопределенные дескрипции. Поскольку количество маркабул увеличилось почти втрое, работу систем машинного обучения можно оценить как стабильную.

Факторы, влияющие на референциальный выбор

Наш подход к моделированию референциального выбора основан на представлении о многофакторности этого процесса (Kibrik, 1996, 1999, 2011). Нашей задачей был не только поиск и изучение факторов, влияющих на РВ, но также изучение их индивидуального вклада в аккуратность предсказания РВ, с целью уменьшения их количества до необходимого минимума (чтобы снизить трудозатратность процесса аннотации). Полный набор факторов включает себя различные характеристики как анафора

и антецедента, так и самого референта, а также некоторые общие дискурсивные характеристики.

- Признаки референта: одушевленность, протагонизм (значимость референта в дискурсе), род и число;
- Признаки антецедента: входит ли в состав прямой речи, тип синтаксической группы, грамматическая роль, референциальная форма, длина антецедента в словах, количество антецедентов в цепочке от текущего места до полной именной группы
- Признаки анафора: первое/непервое упоминание в дискурсе, входит ли в состав прямой речи, тип синтаксической группы, грамматическая роль, число упоминаний референта в цепочке
- Расстояния между анафором и антецедентом: линейное расстояниев словах, линейное расстояние в клаузах, линейное расстояние в предложениях, расстояние в маркабулах, риторическое расстояние в ЭДЕ, расстояние в абзацах.

Список параметров периодически пополняется. Например, одно из последних дополнений к нему связано с исследованием влияния жанрового разнообразия корпуса на РВ. Этот вопрос неоднократно поднимался в предыдущих исследованиях РВ (Biber, Johansson, Leech, Conrad & Finegan, 1999: 235; Fox, 1987; Garrod, 2011; Longo & Todirasu, 2010; Toole, 1996; Strube & Wolters, 2000, De Clercq et al. 2011). Исследователи пришли к выводу, что жанр дискурса влияет на референцию. Несмотря на относительную однородность корпуса RefRhet2, в нем можно выделить три основных жанра — это наиболее часто встречающиеся в корпусе тексты (1) информационные заметки и отчеты (корпоративные и финансовые новости), (2) биографические очерки и (3) аналитические статьи и рецензии. Принадлежность к одному из трех жанров была размечена на подкорпусе и использовалась в качестве одного из параметров машинного обучения. Результатом стало повышение аккуратности около 0.5 %–1 % в зависимости от использованного алгоритма. Вклад этого параметра может показаться несущественным, однако, есть несколько причин включить его в полный набор при тренировке алгоритмов машинного обучения. Трудозатратность разметки жанровой принадлежности не высока, улучшение в аккуратности предсказания, которое она дает, сравнимо с вкладом других отдельно взятых факторов, и, что самое важное, это улучшение нельзя проигнорировать или объяснить прочими характеристиками тестового материала.

Некоторые из параметров, используемых нами при моделировании, делают процесс аннотации трудоемким и времязатратным — к примеру, риторическое расстояние. Наличие разметки риторической структуры в корпусе послужило причиной выбора его для данного исследования, кроме того, в наших предыдущих работах уже обсуждался вклад отдельных факторов в общее значение аккуратности (Loukachevitch et al., 2011). Однако так ли необходима разметка по ТРС для конечного результата? Исключение риторического расстояния из списка факторов дает ухудшение в качестве

предсказания всего на 0.6 % для двухклассовой и 0.1 % для трехклассовой задачи. Результат созвучен общепринятому мнению о том, что риторическое расстояние не влияет на выбор между именем собственным и определенной дескрипцией.

Чтобы оценить необходимость приложения дополнительных усилий на этапе аннотации, мы решили элиминировать из исходного набора параметров те, аннотация которых требует существенно больше усилий. Результаты работы системы (с использованием логистической регрессии) представлены в Таблице 5.

Табл. 3. Сравнение результатов машинного обучения с использование более и менее «дорогого» набора параметров

	Аккуратность для двухклассовой задачи	**Аккуратность для трехклассовой задачи**
Полный набор параметров	87.9 %	77.6 %
«Дешевый» набор параметров	85.2 %	75.5 %

Наши результаты свидетельствуют об ощутимой (хотя и не критической в целом) потере в аккуратности. Это подтверждает, что ни один из факторов не теряет своей значимости, если мы стремимся достичь наилучшего результата.

Деревья решений — взгляд изнутри

Один из наиболее интересных вопросов относится к задаче референциального выбора между местоимением, именем собственным и определенной дескрипцией. Изначально столь высокий показатель аккуратности был не вполне ожидаем, поскольку эксперименты по выявлению признаков, влияющих на различение говорящим имени собственного и определенной дескрипции, дали достаточно скромный результат (Linnik 2010). В настоящем исследовании мы предприняли попытку пойти «обратным путем» и разобрать детально результат работы одного из алгоритмов машинного обучения — деревьев решений. В Таблице 4 представлены некоторые из сгенерированных системой правил с высокой степенью вероятности предсказания.

Основная тенденция, которую можно вывести из этих правил, заключается в наличии эффекта *референциальной инерции* при выборе между вариантами полной именной группы. А именно, если референт упоминается с помощью полной ИГ, будь то дескрипция или имя собственное, при следующем упоминании с высокой вероятностью также будет использована полная и именная группа (см. Krahmer & Theune, 2002).

Табл. 4. Результат обратной обработки деревьев решений (SD — расстояние в предложениях)

Если	То	Количество правильно предсказанных форм	Процент правильных предсказаний из всех случаев, удовлетворяющих условиям из колонки «если»
Референциальная форма антецедента: не дескрипция SD=1 Одушевленность: collective	имя собственное	290	70%
Референциальная форма антецедента: не дескрипция SD>1 Число: единственное	имя собственное	522	86%
Референциальная форма антецедента: дескрипция без определителя Референциальная форма антецедента: не имя собственное	дескрипция	315	90%
Референциальная форма антецедента: дескрипция с определенным артиклем	дескрипция	360	78%
Референциальная форма антецедента: дескрипция с неопределенным артиклем	дескрипция	72	84%

О категоричности референциального выбора

Даже при использовании полного набора факторов наша система не достигает 100% аккуратности предсказания РВ. Новые параметры, которые добавлялись к исходному набору, также не давали существенного улучшения. С целью выяснить, насколько вообще категоричен выбор, осуществляемый говорящими, была запущена серия экспериментов, первые результаты которых будут опубликованы в ближайшее время (Худякова 2012).

В эксперименте испытуемым предлагалось ответить на вопросы к текстам, на которых ошибалась система машинного обучения (в частности, в первом эксперименте этой серии рассматривались случаи, в которых исходно в тексте было употреблено имя собственное, однако система предсказывала появление местоимения). Половине испытуемых были предложены тексты в их исходном виде; в текстах, предложенных второй половине, имена собственные были заменены на соответствующие местоимения. Оговоримся, что машинное

обучение может давать предсказания с различной степенью вероятности — от 0.5 до 1, что означает, что в одних случаях можно с большей уверенностью ожидать то или иное референциальное выражение, в других же — с меньшей.

Результаты эксперимента показали, что в 95% случаев (из тех текстов, которые не вызвали затруднений в исходном варианте) имя собственное и местоимение взаимозаменяемы, то есть, по всей видимости, равновероятны. Это позволяет сделать вывод, что референциальный выбор не категоричен, и этот вывод подтверждается работой нашей системы машинного обучения.

Заключение

Исследования по изучению факторов, влияющих на референциальный выбор, продолжаются. Мы стремились оптимизировать набор используемых в машинном обучении параметров и аннотационную схему без потери в аккуратности предсказания референциального средства. Нам удалось показать, что существует возможность использовать сокращенный список факторов за счет исключения некоторых параметров, сложных в аннотации. Хотя потеря в аккуратности не критична, очевидно, что лучший показатель достигается использованием наиболее полного набора факторов. Мы также продолжаем тестировать различные признаки референта и дискурса в целом — например, его жанра, — с целью улучшить данный показатель. Основная гипотеза, эксперименты по проверке которой уже запущены, заключается в том, что при среднем уровне активации референта референциальный выбор может быть некатегорическим. Полученные на данный момент результаты свидетельствуют о том, что в большинстве случаев, когда используемая нами система машинного обучения ошибается, замена референциального средства не создает говорящим проблем восприятия. Это свидетельствует о том, что в при среднем уровне активации референта выбор той или иной опции действительно может быть равновероятен.

В будущем мы намерены продолжить данную серию экспериментов. Кроме того, по мере того как объем корпуса RefRhet2 будет расти, планируется провести оценку согласия аннотаторов (Artstein & Poesio 2008), что, возможно, также даст нам некоторое представление о том, как влияет качество аннотации на наши результаты, а также о более сложных случаях неоднозначности референциального выбора.

Литература

1. *Artstein, R., & Poesio, M.* (2008), Inter-coder agreement for computational linguistics (surveyarticle). Computational Linguistics, no. 34(4), pp. 555–596.
2. *Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E.* (1999), Longman grammar of spoken and written English. Pearson Education limited, Harlow.
3. *Carlson, L., Marcu, D., & Okurowski, M. E.* (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In J. van Kuppevelt &

R. Smith (Eds.), Current directions in discourse and dialogue. Kluwer, Dordrecht, pp. 85–112
4. *Fox, B. A.* (1987), Discourse structure and anaphora in written and conversational English. Cambridge, Cambridge University Press.
5. *Garrod, S. C.* (2011), Referential processing in monologue and dialogue with and without access to real world referents. In E. Gibson & N. J. Pearlmutter (Eds.), The processing and acquisition of reference (pp. 273–294). MIT Press, Cambridge, MA.
6. *De Clercq O., Hoste V., & Hendrickx I.* (2011), Cross-domain dutch coreference resolution. In Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011), pp. 186–193.
7. *Hobbs, J. R.* (1985), On the Coherence and Structure of Discourse (Report No. CSLI-85–37). Center for the Study of Language and Information, Stanford University.
8. *Joshi, A. K., Prasad, R., & Miltsakaki, E.* (2006), Anaphora resolution: Centering theory approach. In K. Brown (Ed.), Encyclopedia of language and linguistics. Elsevier, Oxford, pp. 223–230.
9. *Kehler, A.* (2002). Coherence, reference, and the theory of grammar. CSLI Publications, Stanford.
10. *Kibrik A. A.* (1996) Anaphora in Russian narrative discourse: A cognitive calculative account In B. Fox (ed.) Studies in anaphora. Benjamins, Amsterdam, pp. 255–304.
11. *Kibrik A. A.* (1999) Reference and working memory: Cognitive inferences from discourse observation In van Hoek, K., Kibrik A. A. & L. Noordman (eds.) Discourse studies in cognitive linguistics. Benjamins, Amsterdam, pp. 29–52.
12. *Kibrik, A. A.* (2011). Reference in discourse. Oxford University Press, Oxford.
13. *Kibrik, A. A., Dobrov, G. B., Zalmanov, D. A., Linnik, A. S., & Loukachevitch N. V.* (2010) Referencial'nyj vybor kak mnogofaktornyj verojatnostnyj process [Referential choice as a multi-factor probabilistic process]. In A. E. Kibrik (Ed.), Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2010" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010"]. Bekasovo, Moscow region, pp. 173–181. Эта работа по-русски, неясно почему она дана в транслите.
14. *Khudyakova M. V.* (2012) Akkuratnost' modelirovanoja referencial'nogo vybora: ocenka chitateljami [Accuracy of referential choice modeling: reader evaluation]. To be presented at The Fifth International Conference on Cognitive Science, to be held in Kaliningrad, Russia. Кириллица!
15. *Krahmer, E., & Theune, M.* (2002). Efficient context-sensitive generation of referring expressions. In K. van Deemter & R. Kibble (Eds.), Information Sharing: Reference and Presupposition in Language Generation and Interpretation , pp. 223–264. CSLI Publications, Stanford.
16. *Krasavina, O. N. & Chiarcos, Ch.* (2007), PoCoS — Potsdam Coreference Scheme. In Proceedings of the Conference of the Association for Computational Linguistics (LAW). Prague, Czech Republic, pp. 156–163.

17. *Longo, L.,& Todirascu, A.* (2010). Genre-based Reference Chains Identification for French. Investigationes Linguisticae, no. 21, pp. 57–75.
18. *Linnik A. S.* (2010) Linguistic support for computational analysis of the corpus of texts, annotated with respect to the referential theory. Graduation thesis at the Moscow State University, Moscow.
19. *Loukachevitch, N. V., Dobrov, G. B., Kibrik, A. A., Khudyakova, M. V., & Linnik, A. S.* (2011). Factors of referential choice: computational modeling. In A. E. Kibrik (Ed.), Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2010" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010"]. Bekasovo, Moscow region, pp. 518–528.
20. *Mann, W. C. & Thompson, S. A.* (1987), Rhetorical Structure Theory: Toward a functional theory of text organization. Text 8(3), pp. 243–281.
21. *Miltsakaki, E., Prasad, R., Joshi, A. & Webber, B.* (2004), The Penn Discourse Treebank. In Proceedings of LREC 2004, Lisbon, Portugal, pp. 2237–2240.
22. *Strube, M., & Wolters, M.* (2000). A Probabilistic genre-independent model of pronominalization. In Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics. Seattle, pp. 18–25.
23. *Taboada, M., & Mann, W. C.* (2006), Rhetorical structure theory : Looking back and moving ahead. In Discourse Studies, no. 8, pp. 423–459.
24. *Toole, J.* (1996). The effect of genre on referential choice. In T. Fretheim & J. K. Gundel (Eds.) Reference and referent accessibility. Benjamins, Amsterdam, pp. 263–290.
25. *Wolf, F., & Gibson, E.* (2003), Representing discourse coherence: A corpus-based study. In Computational Linguistics, no. 31(2), pp. 249–287.

ПРИЛАГАТЕЛЬНЫЕ *ТЯЖЕЛЫЙ* И *ЛЕГКИЙ* В ТИПОЛОГИЧЕСКОЙ ПЕРСПЕКТИВЕ

Кюсева М. В. (pegas@nm.ru),
Рыжова Д. А. (daria.ryzhova@mail.ru),
Холкина Л. С. (kholkina@gmail.com)

МГУ имени М. В. Ломоносова, Москва, Россия

Настоящий доклад посвящён русским прилагательным *тяжёлый* и *лёгкий*, рассмотренным на типологическом фоне в целях подготовки материала для типологической базы данных по многозначным качественным прилагательным и наречиям. Рассматриваемые лексемы устроены неожиданно симметрично, но имеется и ряд существенных отличий: прилагательное *тяжёлый* чаще используется в прямом значении, чем прилагательное *лёгкий*; у прилагательного *лёгкий*, напротив, лучше развита система переносных значений. Материал сербского, английского, французского и китайского языков дополняет русский и позволяет выделить в семантической структуре прилагательного ‘легкий’ два случая неполной энантиосемии: оно может приобретать такие компоненты значений как «быстрый» и «медленный», а также развивать как положительные, так и отрицательные коннотации. Прилагательное же ‘тяжелый’, в русском языке исходно антропоцентричное и склонное к обозначению только низкой скорости и только отрицательной оценки, может развивать в других языках положительные переносные значения, в случае если его исходным значением является ‘(объективно) имеющий большой вес’.

Ключевые слова: тяжелый, легкий, признаковая лексика, лексическая типология

Исследование, которое мы представим, посвящено прилагательным *тяжёлый* и *лёгкий*. Однако цели нашей работы не сводятся к описанию семантических структур этих двух русских лексем (тем более, что это уже было превосходно сделано в ряде работ: см. описание прилагательного *тяжёлый* в Кустова 2004, а также словарные статьи в НОСС (Урысон 2004) и в Проспекте активного словаря русского языка (Бабаева 2010). Мы рассмотрим эти прилагательные в типологической перспективе и постараемся решить две задачи: теоретическую и практическую.

Практическая задача проста: она заключается в подготовке материала для типологической базы данных по многозначным качественным прилагательным и наречиям. Этот проект является продолжением работы по созданию Базы данных по многозначным качественным прилагательным и наречиям русского языка, который мы подробно представляли на Диалоге-2010 (см. Карпова и др. 2010). Основной его целью было составление максимально полного каталога семантических переходов в русской признаковой лексике, для чего

в Базу данных были занесены семантические структуры около 300 наиболее частотных (по частотному словарю Ляшевской и Шарова, см. Ляшевская, Шаров 2008) русских признаковых слов. В типологическую базу данных по многозначным качественным прилагательным и наречиям планируется включить около 30 признаковых лексем[1], проанализировав их семантические структуры, по крайней мере, в пяти языках. База будет организована по принципу типологической анкеты: для каждого прилагательного будут выделены ситуации (или, в другой терминологии, домены или фреймы, ср., например, Lakoff, Turner 1989, Croft 1993, Blank 1999), которые могут этим прилагательным (или соответствующим наречием) покрываться. Таким образом, наша практическая задача заключается в подготовке «фреймовой анкеты» для прилагательных со значениями ‘тяжёлый’ и ‘лёгкий’, и за основу мы берём родной для нас русский язык, а в качестве фона рассматриваем материалы сербского, французского, английского и китайского языков.

Теоретическая задача связана с практической и подразумевает сравнение семантических структур прилагательных *тяжёлый* и *лёгкий* в разных языках. Для нас важно, что объектом исследования является антонимическая пара лексем, а не два отдельно взятых прилагательных: при анализе семантических структур (и их потенциалов) прилагательных *тяжёлый* и *лёгкий* мы будем одновременно сравнивать эти структуры между собой.

1. Симметрия структур: обманутые ожидания

Обычно даже беглый взгляд на употребление антонимичных лексем показывает, что они антонимичны лишь частично: их семантические структуры симметричны только в некоторых значениях. Например, антонимом к прилагательному *старый*, описывающему возраст человека, животного или растения (ср. *старая женщина/собака/яблоня*), является прилагательное *молодой* (ср. *молодая женщина/собака/яблоня*), а антонимом к прилагательному *старый*, обозначающему «возраст» артефактов (ср. *старый шкаф*), служит прилагательное *новый* (ср. *новый шкаф*). *Тяжелый* и *легкий*, как кажется на первый взгляд, нарушают эту закономерность. Помимо прямого употребления (*тяжелый чемодан — легкий чемодан*), они обнаруживают структурное сходство в значениях ‘трудный / простой’ (*тяжелая задача — легкая задача*), ‘серьезный, глубокий / веселый, имеющий развлекательный характер’ (*тяжелый фильм — легкая комедия*), ‘неприятный /приятный в общении’ (*тяжелый ребенок — легкий человек)*, ‘мрачный, безрадостный / свободный, беспечный’ (*тяжелая жизнь — легкая жизнь*), ‘перевариваемый, усваиваемый с трудом/без труда’ (*тяжелая пища — легкая пища*) и в некоторых других. Единственное отличие, которое

[1] Как показывает практика нашей работы с русскими прилагательными, наиболее разветвленные сети значений имеют частотные слова, первым значением которых является обозначение физического свойства предмета, такие как *мягкий, горячий, острый, ровный* и т. п.

удаётся обнаружить в результате простого составления списков значений одной и другой лексемы, — это способность прилагательного *легкий* выражать значение ‘тонкий, не стесняющий движений’ (*легкая одежда*) и отсутствие этой возможности у прилагательного *тяжёлый (ср. теплая одежда)*.

До сих пор мы рассматривали симметрию прямых и переносных значений изучаемых слов в пределах одного языка — ее можно назвать «внутренней». Однако «нетрадиционность» прилагательных *легкий* и *тяжелый* проявляется и в другой их характерной черте — симметрии «внешней». Для лексических типологов привычна ситуация, когда одной лексеме в русском языке соответствуют несколько слов в некоторых других языках (или наоборот). Причина состоит в том, что в одном языке лексема в своем первом значении может покрывать больший пласт «ситуаций» (= большее количество фреймов), чем эквивалентная ей лексема в другом. Например, русскому прилагательному *острый* в коми-зырянском языке соответствуют прилагательные *лэчыд* и *ёсь*: первое описывает инструменты с острым краем (*лэчыд чер* — острый топор), а второе — с острым кончиком (*ёсь штык* — острый штык). К нашему удивлению, ‘легкий’ и ‘тяжелый’ ни в одном из рассмотренных нами индоевропейских языков не «разделились»: в английском, во французском и в сербском языках русским прилагательным соответствуют две антонимичные лексемы (*heavy* & *light*, *lourd* & *léger*, *težak* & *lak,*), причём устроены они примерно так же симметрично, как русские *тяжёлый* и *лёгкий*.

2. Асимметрия: на круги своя

Кажущаяся симметрия нарушается при рассмотрении материала китайского языка — одному ‘легкому’ (*qīng)* в нем соответствуют два ‘тяжелых’ *(zhòng* и *chén*)[2]. Это подтвердило наши исходные предположения о различной организации этих признаковых полей, и мы сосредоточили свое внимание на проявлениях асимметрии в семантических структурах рассматриваемых лексем.

Прежде всего, мы обратились к статистике и обнаружили, что русская лексема *лёгкий* в целом частотнее лексемы *тяжёлый* (82 025 вхождений по НКРЯ против 61 072 соответственно), но *тяжёлый* значительно чаще употребляется в своём прямом значении: в НКРЯ нам встретилось 73 примера на сочетание *тяжёлая сумка* и всего 1 на сочетание *лёгкая сумка*; 63 примера сочетания

---

[2] Z*hòng u chén* имеют разные исходные значения. Z*hòng* описывает вес как большой и с объективной, и с субъективной точки зрения и часто развивает переносные значения с положительной оценкой (см. примеры (12)–(14)), Исходное значение c*hén* — «тонуть», что подтверждается и графической формой соответствующего иероглифа — в него входит смысловой детерминатив «вода». В переносных употреблениях он описывает состояния/события как отрицательные (например, *tóu chén* «голова тяжелая», *chén tòng* «тяжелая боль» — сильная боль). В нормативном китайском языке для обоих этих прилагательных постулируется одинаковое значение ‘тяжелый’ (Словарь современного китайского языка Xiandai Hanyu cidian), однако в диалектах они часто оказываются дополнительно распределены.

тяжёлый чемодан и всего 10 — *лёгкий чемодан*. Таким образом, *лёгкий*, в отличие от *тяжёлого*, больше «сдвинут» в сторону переносных значений. Эта гипотеза подтверждается и тем, что у прилагательного *тяжёлый* есть несколько квазисинонимов именно в метафорических контекстах: ср. *тяжкий/тяжёлый грех, трудная/сложная/тяжёлая задача, тягостное/тяжёлое впечатление*, в то время как прилагательное *лёгкий* покрывает свои переносные значения в основном самостоятельно, сильно пересекаясь только с прилагательным *простой*.

Неравномерное распределение рассматриваемых лексем по зонам прямых/переносных значений иллюстрируется и контекстами, предполагающими значение (анти)интенсива (лексическая функция (Anti)Magn в терминологии Мельчука, см. Мельчук 1974). Так, *тяжелый* в некоторых сочетаниях имеет значение ‘сильный, проявляющийся с большой степенью’ (ср., *тяжелое повреждение, тяжелые переживания, тяжелое похмелье)*, но антонимичное значение прилагательного *легкий* развито намного сильнее. Возможны сочетания *легкие морщинки,* но не **тяжелые морщины, легкий ветерок*, но не **тяжелый ветер, легкий смешок*, но не **тяжелый смех, легкий румянец,* но не **тяжелый румянец*. В качестве антонимов прилагательного *лёгкий* в этих контекстах используются лексемы *глубокий, сильный, громкий* и *яркий* соответственно.

В других рассмотренных нами языках это различие не столь яркое. Сербский, английский, французский и китайский языки демонстрируют большую развитость значения интенсива у прилагательного ‘тяжелый’ по сравнению с русским коррелятом. Так, в английском языке возможны сочетания *heavy cold* (букв. «тяжелый насморк»), *heavy thunderstorm* (букв. «тяжелая гроза»), *heavy rain* (букв. «тяжелый дождь»). Во французском возможно редкое для русского языка сочетание *chaleur lourde* (букв. «тяжелая жара»). Нужно сказать, что поле «(анти)интенсивности» у прилагательных ‘тяжелый’ и ‘легкий’ в этих языках вообще развито лучше, чем в русском. Одно из доказательств тому — расширение количества допустимых контекстов для этого значения до такой степени, что появляются слова, в сочетании с которыми рассматриваемые прилагательные начинают передавать смежный смысл — ‘большой по количеству’. Так, во французском языке распространены сочетания *lourdes dépences* (букв. «тяжелые расходы»), *lourdes dettes* (букв. «тяжелые долги»), *frais lourds* (букв.: «тяжелые издержки»). В китайском языке ‘тяжелой’ может быть ‘цена’ (*zhòng jià*) или ‘награда’ (*zhòng jiǎng*). Все эти контексты предполагают идею ‘много денег’. Заметим, что в русском языке подобные сочетания возможны (ср., *тяжелая дань*), однако по большей части они устарели и употребляются достаточно редко.

Интересен также сербский пример *teške pare* (букв. «тяжелые деньги» = много денег). В русском языке его коррелят обладает совсем другим значением: *тяжелые деньги* — это не большие деньги, а деньги, которые добывались тяжело, с большим трудом.

Идея количества проявляется и в употреблениях прилагательного ‘легкий’, ср.:

(1) *Un budget trop **léger*** — слишком **маленький** бюджет (фр.)

(2) *niánjì hěn* ***qīng*** — молодой (букв.: «возраст очень лёгкий») (кит.)

Таким образом, благодаря материалам английского, французского, сербского и китайского языков в типологической анкете появляется новый фрейм, связанный с идеей количества.

Второе важное поле, демонстрирующее асимметрию прилагательных ‘тяжелый’ и ‘легкий’ — поле скорости. Русское прилагательное *тяжелый* в некоторых контекстах приобретает компонент значения ‘медленный’. Такая коннотация логично связана с первым значением прилагательного: движение нагруженного человека, животного или транспортного средства обычно медленнее, чем движение без груза (подробнее об этом см. уже упоминавшуюся выше работу Кустова 2004). Таким образом, *тяжелая походка* — медленная, грузная походка, *тяжелые движения* — медленные движения. Аналогичные сочетания встречаются в сербском языке, ср. *čovek teških kretanja* (букв. «человек с тяжелыми движениями»), а также *čovek teški od vina* (букв. «человек, тяжелый от вина» = человек, который медленно и с трудом передвигается из-за того, что выпил вина). Проявляется компонент ‘медленный’ и во французском прилагательном *lourd,* ср.:

(3) *Un équipement qui gêne le mouvement, qui rend* ***lourd*** *et maladroit*. — Обмундирование, которое стесняет движение, которое затрудняет и **замедляет** ход.

Прилагательное *легкий* может приобретать противоположный компонент значения — ‘быстрый’: если человек идёт *налегке*, не нагруженный тяжелыми вещами, скорость его передвижения увеличивается. Ср., например:

(4) *Быстрой и* ***легкой походкой*** *Ярый пересек пару узких улочек…* [Андрей Грачев. Ярый-3. Ордер на смерть (2000)][3]

Сочетание *легкие войска* также обозначает войска, не нагруженные тяжелыми орудиями и, как следствие, имеющие возможность быстро передвигаться[4].

Однако у прилагательного ‘легкий’ есть и противоположное значение, синонимичное значению ‘тяжелого’, — ‘медленный’. Ср., например, русские предложения:

(5) *Ребята потопали* ***легкой трусцой*** *к своим, идущим впереди, метрах в тридцати — сорока.* [Захар Прилепин. Санькя (2006)]

(6) *В тридцати минутах* ***легкого бега*** *от дачи Воротникова было несколько прудов.* [Петр Акимов. Плата за страх (2000)]

---

[3] Здесь и далее для русского материала приводятся примеры из НКРЯ.

[4] В древнерусских текстах XVII в. встречаются также сочетания ‘лёгкая лошадь’, ‘лёгкий зверь’ в значении «быстрая лошадь», «быстрый зверь» (см. Лукина 1962)

Здесь *легкий* означает ‘спокойный, небыстрый’. Похожий смысл передает и наречие *легонько*, а также слово *легко* в сравнительной степени — *полегче*.

(7) — *Ты, Михайло Антонов, не торопись, хорошо — скоро не бывает.* ***Легонько*** *надо!* [Максим Горький. Мои университеты (1923)]

(8) — *И* ***полегче*** *на поворотах, — неожиданно попросил Алексей Александрович.* [Юрий Герман. Дорогой мой человек (1961)]

Крайняя степень развития этого значения наблюдается у сербского прилагательного *lagani* (аттенуатив от *lak*, коррелят русского *легонький*) и соответствующего наречия *lagano* (ср. русское *легонько*). В большей части своих употреблений оно выражает именно значение ‘медленный’:

(9) *Uputili su se* ***laganim korakom*** *prema jednoj vili gotovo na samoj obali.* — Они отправились **медленным шагом** в сторону виллы, находящейся почти на самом берегу.

Возможное объяснение такой внутренней антонимии лексем со значением ‘легкий’ заключается в том, что компонент ‘медленный’ естественным образом вытекает из развитого значения «анти-интенсива». Действительно, в сочетании с некоторыми абстрактными сущностями значение слабой степени заключается именно в низкой скорости, ср. *лёгкий бег*: «слабая степень» бега.

Наконец, третий компонент, который демонстрирует ассиметрию ‘легкого’ и ‘тяжелого’ — это оценка. С русским прилагательным *тяжелый* логично связана отрицательная коннотация, поскольку в своем первом значении оно обозначает переизбыток веса, доставляющий неудобство. *Тяжелая сумка* — это не просто сумка, в которой много вещей, это сумка, нести которую *тяжело*. Отрицательная оценка сохраняется и в метафорических значениях: *тяжелая жизнь* — это жизнь трудная, полная проблем; *тяжелый разговор* никогда не бывает приятным; *тяжелые известия* всегда печальны; с *тяжелым человеком* неприятно общаться, потому что его характер сложно *вынести*. Словосочетание ‘тяжелый человек’ возникает и в других рассмотренных нами языках, ср. *heavy man, homme lourd*. Здесь отрицательный компонент также играет существенную роль, однако в целом прилагательное ‘тяжелый’ приобретает другое значение: грубоватый, глуповатый, медленно соображающий, ср.:

(10) *If there is anything worse it is* ***a heavy man*** *when he fancies he is being facetious.* — Если и бывает что-нибудь хуже, так это **глупый человек**, который считает себя остроумным.

Во французском языке это значение также дает метонимический перенос на описание текстов и устного дискурса:

(11) *Mais il ne participe pas aux* ***lourdes plaisanteries*** *de ses collègues.* — Но он не участвует в **глупых шутках** своих коллег.

Напротив, если ноша легка и не доставляет неудобств, это приятно. Таким образом, положительные коннотации, связанные с прилагательным *легкий* вполне ожидаемы. С *легким человеком* приятно общаться, *легкая комедия* смотрится с удовольствием и без напряжения, *легкая жизнь* — это жизнь без забот и лишений. Подобные словосочетания встречаются и в других языках, ср. французское *comedie légère* (легкая комедия) и английское *light reading* (легкое чтение). А китайское сочетание *wú bìng yì shēn qīng* (букв. «нет-болезнь-один-тело-легкий»)обозначает «нет болезней — все тело легкое» (ср. русское *легкость во всем теле*).

Однако и в данном случае ‘легкий’ демонстрирует энантиосемию: с ним может быть связана и отрицательная оценка. *Tête légère* (букв. «легкая голова») во французском языке — это голова пустая, *homme léger* — это несерьезный, беспечный, безответственный человек, а *femme légère* — женщина чересчур свободной манеры поведения (ср. русское *женщина лёгкого поведения*). Аналогичное сочетание есть и в английском языке: *light woman* — кокетка. В русском языке для передачи подобных значений используется прилагательное *легкомысленный*. Чем же может быть мотивирована эта негативная оценка? Как кажется, дело в том, что ‘легкий’ значит «без излишнего груза», но ведь груз может быть и ценным.

Именно с этой идеей связан еще один важный пласт переносных значений, которые ‘легкий’ и ‘тяжелый’ развивают в некоторых из исследуемых нами языков. Рассмотрим примеры китайского и сербского языков:

(12) *Qīng* ***zhòng*** *dǎo zhì* — Не разбираться, что **важно**, что неважно (букв. «легкий-тяжелый-переворачивать-ставить»)

(13) *Wǒmen de rènwu hěn* ***zhòng*** — Перед нами стоят очень **важные** задачи (букв. «мы-ATR-задача-очень-тяжелый»)

(14) *Nǐ zài wǒ xīn li fènliàng hěn* ***zhòng*** — Ты для меня очень много **значишь** (букв. «ты-в-я-сердце-внутри-вес-очень-тяжелый»).

(15) *Crkovna je vlast* ***teža*** *od svjetovne*. Церковная власть **важнее** (=весомее, букв.: «тяжелее») светской. (Словарь сербского языка Матица Српска)

Во всех этих примерах ‘тяжелый’ выражает значение ‘важный, весомый, значимый’.

На первый взгляд, это противоречит нашему предположению о том, что ‘тяжелый’ развивает только отрицательные коннотации. Однако при ближайшем рассмотрении становится ясно, что речь идет о двух разных ‘тяжелых’ — антропоцентричном ‘тяжёлом’, описывающем предметы, которые *человеку тяжело нести, поднимать* (ср. связь внутренней формы данного прилагательного

с глаголом *тянуть*) (см. Кустова 2004) и объективным ‘имеющим большой вес’ (ср. *весомый, веский*).

Действительно, и в сербском, и в китайском вопрос об объективном весе (Сколько ты весишь?) задается с использованием прилагательного ‘тяжёлый’:

(16) *Koliko si* ***težak****?*

(17) *Nǐ yǒu duō* ***zhòng****?* (букв. «ты-иметь-насколько-тяжелый»).

Таким образом, мы выявили ещё один важный параметр, релевантный для лексем со значением ‘тяжёлый’: антропоцентричность vs. объективность. Необходимо отметить, что данный параметр уже упоминался в работах Урысон (Урысон 2004) и Бабаевой (Бабаева 2010), однако постулировался как релевантный для семантической структуры русской лексемы *тяжёлый*. На наш взгляд, этот параметр значим для концепта ‘тяжёлый’ в целом, но не для конкретного русского прилагательного *тяжёлый*, которое, с нашей точки зрения, антропоцентрично.

Итак, анализ прилагательных со значением ‘тяжелый’ и ‘легкий’ в русском языке с привлечением в качестве фона данных сербского, английского, французского и китайского языков показывает, что эти поля устроены хоть и сходно, но не одинаково. Различия наиболее ярко проявляются в зонах интенсификации, скорости и оценки. Так, у русского прилагательного *лёгкий* значительно больше развито значение анти-интенсификатора, чем у прилагательного *тяжёлый* — значение интенсификатора, в то время как в других рассмотренных нами языках эта зона в целом развита сильнее и включает помимо значения высокой/низкой степени ещё и значение количества. Кроме того, проанализированный материал позволяет выделить в семантической структуре концепта ‘лёгкий’ два случая неполной энантиосемии, которые релевантны и для русского языка, но ярче видны при привлечении к исследованию типологического фона. Прилагательное со значением ‘лёгкий’ может приобретать такие компоненты значений, как «быстрый» и «медленный», а также развивать как положительные, так и отрицательные коннотации. Напротив, русское прилагательное *тяжёлый* склонно к обозначению только низкой скорости и только отрицательной оценки, что связано с его антропоцентричностью — изначально оно описывает предметы, которые *человеку тяжело нести, поднимать*. Типологически оно также может развивать неполную энантиосемию с точки зрения оценочного компонента, но только в том случае, если оно изначально способно выступать не только в отрицательных (антропоцентричных), но и в нейтральных контекстах[5].

[5] Как показывают наши данные, случаи энантиосемии у прилагательных со значением ‘тяжёлый’ встречаются по языкам заметно реже, чем случаи энантиосемии в прилагательных со значением ‘лёгкий’. Однако наша выборка ещё слишком мала для того, чтобы можно было уверенно делать такого рода обобщения.

Выявленные особенности прилагательных данного семантического поля, прежде всего, нарушения как «внутренней», так и «внешней» симметрии, послужат основой «фреймовой анкеты» для типологической базы данных по многозначным качественным прилагательным.

Литература

1. *Бабаева Е. Э.* Антонимия: проблемы толкования и реконструкции становления (на примере прилагательных с сильно развитой многозначностью) // акад. Апресян Ю. Д. (ред.). Проспект активного словаря русского языка. М.: Языки славянских культур, 2010.
2. *Карпова О. С., Архангельский Т. А., Кюсева М. В., Рахилина Е. В., Резникова Т. И., Рыжова Д. А., Тагабилева М. Г.* База данных по многозначным качественным прилагательным и наречиям русского языка // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2010» (Бекасово, 26–30 мая 2010 г.). Вып. 9 (16).– М.: РГГУ, 2010.
3. *Кустова Г. И.* Типы производных значений и механизмы языкового расширения. М.: Языки славянской культуры, 2004
4. *Лукина Г. Н.* Антонимические прилагательные в памятниках древнерусского языка // Р. И. Аванесов (ред.). Историческая грамматика и лексикология русского языка, материалы и исследования. М.: изд-во Академии наук СССР, 1962.
5. *Ляшевская О. Н., Шаров С. А.* Новый частотный словарь русской лексики. 2008. http://dict.ruslang.ru/freq.php
6. *Мельчук И. А.* Опыт теории лингвистических моделей «Смысл ↔ Текст». М., 1974.
7. *Речник* српскохрватскога књижевног језика «Матица Српска» [Словарь сербохорватского литературного языка «Матица Српска»]. Нови Сад — Загреб, 1967.
8. *Словарь* современного китайского языка — «Xiandai Hanyu cidian». — Пекин: Коммерческая пресса, 2005.
9. *Урысон Е. В.* Тяжёлый 1.1, увесистый 1, неподъёмный // Новый объяснительный словарь синонимов русского языка. Второе издание, исправленное и дополненное. Под общим руководством акад. Ю. Д. Апресяна. Москва; Вена: Языки славянской культуры: Венский славистический альманах, 2004.
10. *Blank, A.* Co-presence and Succession: A Cognitive Typology of Metonymy // K.-U. Panther, G. Radden (eds). Metonymy in Language and Thought. Amsterdam/ Philadelphia: John Benjamins, 1999
11. *Croft, W.* Case marking and the semantics of mental verbs // J. Pustejovsky (ed.) Semantics and the lexicon. — Dordrecht: Kluwer Acad. Publ., 1993
12. *Lakoff, G.; Turner, M.* 1989. More than cool reason: A field guide to poetic metaphor. Chicago: Univ. of Chicago

БАЗА ДАННЫХ «ТЕЛО И ТЕЛЕСНОСТЬ В ЯЗЫКЕ И КУЛЬТУРЕ (ИДЕОЛОГИЯ, СТРУКТУРА И НАПОЛНЕНИЕ)»

Клыгина Е. А. (katerinakk@rambler.ru),
Крейдлин Г. Е. (gekr@iitp.ru)

Институт лингвистики РГГУ, Москва, Россия

В статье представлена компьютерная вопросно-ответная система, которая отвечает на вопросы, относящиеся к семиотической концептуализации тела. Семиотическая концептуализация тела — это понятие, которое призвано отразить типичные взгляды неискушенных носителей разных языков и культур на человеческое тело и которое учитывает взаимодействие в устном диалоге вербального и невербального знаковых кодов.
Основу данной системы составляет база данных, содержащая разного рода сведения о телесных, или соматических, объектах, об их структурных, физических и функциональных признаках и о действиях, которые эти объекты могут выполнять или которые могут выполняться над ними. Кроме того, система устанавливает соответствия между признаками, значениями признаков и их типовыми языковыми выражениями. В статье рассматриваются внутренняя структура описываемой базы данных и ее формальные и содержательные ресурсы, которые обеспечивают интерфейс между пользователем и компьютером.

Ключевые слова: вопросно-ответная система, тело, телесность, культура, знаковый код, естественный язык, язык тела, соматический (телесный) объект, признак, значение признака

1. Постановка задачи. Понятие семиотической концептуализации тела

В настоящей работе представлены общая идеология, структура и содержательное наполнение компьютерной базы данных «Тело и телесность в культуре и языке», которая создана в Институте лингвистики РГГУ (Москва) небольшим коллективом людей под руководством проф. Г. Е. Крейдлина[1]. Эта база данных

[1] В исследовательскую группу входят преподаватели и аспиранты РГГУ: проф., д. фил. н. Г. Е. Крейдлин, канд. фил. н. П. М. Аркадьев, канд. фил. н. А. Б. Летучий, А. Г. Кадыкова, Э. Е. Заришева, Е. А. Клыгина, Л. А. Хесед и С. И. Переверзева, а также некоторые студенты, привлекаемые к работе над отдельными частями проекта.

является результатом работы над двумя связанными между собой научно-исследовательскими проектами[2].

Первый из них — «Части тела в русском языке и русской культуре» — был закончен в 2010 году, и тогда же Э. Е. Заришевой и Е. А. Клыгиной был разработан первоначальный макет базы. При работе над вторым проектом — «Тело, его части в разных языках и культурах: типологическое описание» — выявилась необходимость корректировки и дополнения инструментария базы с целью приспособить ее к типологическим изысканиям.

Прежде чем приступить к анализу структуры и содержания базы данных и характеристике ее технического обеспечения, которые составляют основу справочного портала о теле, частях тела и других телесных, или соматических, объектах (далее — СО)[3], обратимся к важному теоретическому понятию **семиотической концептуализации тела**. Оно было ранее введено в серии работ участников проектов[4]. Здесь же мы лишь уточним некоторые аспекты этого понятия.

Специалисты в области лингвистики, невербальной семиотики и искусственного интеллекта в отличие, например, от представителей естественных наук изучают тело человека не как физический или биологический объект. Их интересует наивная семиотическая концептуализация тела.

Понятие «семиотическая концептуализация фрагмента мира» является естественным расширением понятия **языковой концептуализации <**данного фрагмента>. Необходимость его введения вызвана тем, что в актах устной коммуникации людей используется не только естественный язык, но другие знаковые (семиотические) коды. Из последних нас будет интересовать телесный код, лексику которого образуют знаковые движения человеческого тела. Лексические единицы телесного кода вместе с грамматикой, то есть правилами соединения и упорядочивания этих единиц, образуют так называемый язык тела (**body language**). Наряду с естественным языком знаки, категории и другие аспекты языка тела определяют характер и течение коммуникации, а также содержание решаемых в ее рамках задач.

Семиотическая концептуализация такого фрагмента мира, как человеческое тело, — это отражение в сознании наивного человека того, как различные СО представлены в рассматриваемых нами знаковых кодах. Мы говорим здесь *наивного* потому, что нас интересуют взгляды и мысли о теле и телесном поведении обычного, так сказать, наивного, человека, а не искушенного специалиста-ученого.

В ходе построения семиотической концептуализации тела мы не только изучаем, как тело представлено в двух семиотических кодах, но и сравниваем

---

[2] Оба проекта в разные годы были поддержаны грантами РГНФ: (РГНФ, грант № 070400203А, РГНФ, грант № 100400125А). Настоящая работа выполнена в рамках второго проекта.

[3] К СО мы относим тело и самые разные его части, в том числе органы, телесные покровы, жидкости, текущие в теле, кости, мышцы и т. д.

[4] О понятии «семиотическая концептуализация тела» см. Крейдлин, Переверзева 2010.

выразительные возможности кодов, т. е. проводим сопоставительное описание естественного (русского) языка и соответствующего ему (русского) языка тела. Существенно, что обе задачи, а тем самым и построение рассматриваемой здесь базы данных, решаются в рамках единого лексикографического формата и единого метаязыка, который создается в ходе их решения.

Подход к описанию тела и телесности, принятый в наших проектах, мы назвали *признаковым*. Его применимость к решению поставленных в проекте задач была теоретически обоснована и эмпирически проверена нами на сравнительно большом вербальном и невербальном материале. В рамках признакового подхода осуществляется описание ядерной части семиотической концептуализации тела, которую составляют десять множеств. Среди них множество СО и их имен[5], таких как *щеки*, *спина*, *ноги*, *живот*, *пальцы*, *язык*, *пупок* и др.[6], множество признаков СО («форма», «ориентация», «цвет», «функция», «температура» <тела и других СО> и др.[7]) и множество жестов и классов жестов (в проектах были изучены как отдельные жесты, так и классы жестов, например, объятия и поцелуи[8]).

Скажем буквально несколько слов еще об одном множестве — множестве значений признака СО. Из всех возможных значений признака нас интересуют главным образом **выделенные** (**salient**) **значения** признака. Выделенность понимается либо как **семантическая выделенность** (это такая характеристика объекта, которая говорит нам что-то существенное о человеке — его обладателе), либо как **культурная выделенность** (последняя говорит нам что-то о той культуре, которой принадлежит данный человек или с которой он тесно связан). Приведем несколько примеров.

Семантически выделенным значением признака «размер головы» является /большой/, так как по меньшей мере два языковых выражения этого значения — прилагательное *большой* в сочетании *большая голова* и сложное прилагательное *большеголовый* — говорят не только о соответствующей величине головы, но и об уме человека.

Культурно выделенными являются, например, некоторые значения классификационного признака — «тип СО». Так, среди японских поз есть тип «позы богов», а внутри него есть поза *агуми маситэ* «сидеть на острие меча, скрестив ноги». По японским преданиям боги ставили меч острием вверх и садились

---

[5] Об этих множествах вместе с примерами их описаний см. статью Крейдлин, Переверзева 2011, а также другие статьи участников проекта.

[6] Здесь и далее в работе приняты следующие условные обозначения: СО специально никак не выделяются, а их имена выделяются курсивом. Признаки СО заключаются в двойные угловые кавычки «», а значения признаков обрамляются одинарными косыми черточками //. Смыслы языковых и неязыковых знаков заключаются в марровские кавычки, или лапки ''.

[7] См., в частности, работы Аркадьев, Крейдлин 2008, Аркадьев, Крейдлин 2011, Кадыкова, Крейдлин 2010, Крейдлин, Летучий 2006, Крейдлин, Летучий 2010, Крейдлин, Переверзева 2008.

[8] См. Крейдлин, Переверзева 2012а,б.

на кончик лезвия, демонстрируя этим свою чудесную силу; см. Крейдлин 2002. Таким образом, эта поза является культурно выделенной во множестве поз богов <в японской культуре>. А в русской культуре культурно выделенным значением является значение ‘глубокий’ признака «физическая реализация поклона». Оно характеризует сгибание тела вдоль оси ‘верх-низ’ и используется для выражения смыслов ‘уважение’, ‘почтение’ или ‘смирение’. В русском языке этот поклон может отображаться разными способами, например, при помощи слов *глубокий*, *низкий* и *земной*. Если соответствующие языковые выражения *глубокий поклон* и *низкий поклон* (точнее их переводные эквиваленты) встречаются также и в других культурах, то выражение *земной поклон,* плохо переводимое на другие языки, является характерным именно для русской культуры; оно является в ней культурно выделенным.

2. Назначение и содержательное наполнение базы данных

Представляемая база данных была задумана как компьютерная система, предназначаемая для ответа в автоматическом режиме на ряд вопросов о семиотической концептуализации тела и телесности. Мы стремились к тому, чтобы наша система была **гибкой**, а это означает, во-первых, что в ней хранятся сведения о теле и телесности, применительно ко всем естественным языкам и языкам тела, которые включены в систему, и, во-вторых, что ее сравнительно легко можно приспособить к новым языкам и соответствующим языкам тела.

Информационные вопросы[9], которые разрешается задавать базе данных, делятся на несколько групп.

1. Группа вопросов о СО. Она подразделяется на ряд подгрупп в зависимости от темы вопроса. В принципе, тема может быть достаточно произвольной, и, поскольку в системе хранится вся информация о каждом СО, теоретически мы можем получить ответ на любой заданный о данном объекте вопрос. Между тем мы считаем, что разумно задавать системе только вопросы строго определенной тематики, а именно те, которые сообщают человеку сведения, пополняющие его знания о семиотической концептуализации тела. Например, запрещено задавать системе метаинформационные вопросы типа *Что такое соматический объект?, Каких признаков вообще не бывает у соматических объектов?* или *Какие есть соматические объекты, обозначаемые пятибуквенными словами?*.

Зато система умеет отвечать на вопросы, относящиеся к группе классификационных, в частности, на вопрос *Какие соматические объекты принадлежат хранящимся в системе типам соматических объектов?,* или на серию вопросов об именах СО: *Как называются соматические объекты, принадлежащие имеющимся в систем типам?*. Кроме того, в нашей базе имеются средства, позволяющие отвечать на ряд других содержательных вопросов о СО или об их именах, например *Какие значимые части выделяются в данном языке*

[9] Информационные вопросы мы отличаем от метаинформационных вопросов, см. о них в тексте ниже.

(или: *в соответствующем языке тела*) *у данного соматического объекта?*, *Как называются значимые части…?*. Сюда же примыкает вопрос *Какие соматические объекты принадлежат типу* «части тела»*?*, или, проще, *Какие имеются <в памяти системы> части тела?* Ответ на все подобные вопросы база выдает в виде множества СО {рука, нога, голова...}. В то же время на вопрос *Каковы имена соматических объектов из типа «части тела»?,* или *Как называются части тела?* ответ предполагается в виде множества имен: {*рука*, *ручонка*, *ручище*, *нога*, *ножка*, *голова*, *головушка*…}[10].

Наша система приспособлена и к тому, чтобы отвечать на «противоположные» вопросы — вопросы о принадлежности данного СО тому или иному типу (или на аналогичные вопросы об их именах), ср. *Какому типу соматических объектов принадлежит хрящ?*, *Как называется тип соматических объектов, которому принадлежит кровь?*, *Слово «волосики» — это название соматического объекта какого типа?* или *Каково имя типа* (*класса*) *соматических объектов, к которому относится объект с именем пупок?* Первый из перечисленных вопросов — это вопрос о СО. Второй — об имени класса, в который входит указанный объект (кровь), а третий — об отнесении объекта с данным именем к конкретному типу СО. Наконец, четвертый вопрос — это вопрос об имени объекта.

2. Вторая группа вопросов касается соотношения СО и их признаков. Это вопросы о том, какими признаками обладают те или иные СО (или аналогичный вопрос об именах объектов), а также противоположный вопрос типа *У каких соматических объектов есть данный признак?* (то же — про названия объектов и признаков). Наша система отвечает, в частности, на вопросы *Какими признаками обладает щека?*, *Какие соматические объекты могут издавать звуки?* или *В каких соматических объектах звуки распространяются?*[11].

3. Третью группу составляют вопросы, темой которой являются значения признаков и имена этих значений. Наша система отвечает на вопросы типа *Какие значения имеет признак форма глаз?* или, более общо, *Какие значения может иметь признак форма соматического объекта?*.

Система знает, что значения некоторых признаков делятся на содержательные классы. Например, в работе Аркадьев, Крейдлин 2008 были выделены геометрические и негеометрические значения признака «форма». Так, одним из геометрических значений признака «форма щек» является /круглые/, а одним из негеометрических — /надутые/. Система умеет отвечать на вопросы типа *Какие геометрические* (*негеометрические*) *значения имеет форма щек?*. Точно так же, поскольку система знает, что у глаз бывает постоянный цвет (значения: /карие/, /голубые/, /черные/…) и переменный цвет (значение этого признака: /красный/), она отвечает на вопросы о значениях постоянного и переменного цвета глаз.

В ближайшем будущем мы надеемся обучить систему отвечать на разнообразные вопросы о выделенных значениях признаков, в частности, *Какие*

---

[10] Впрочем, пока система выдает только стандартные имена СО, не учитывая их словообразовательных вариантов.

[11] См. Крейдлин, Переверзева 2010.

значения данного признака являются выделенными?, Являются они семантически или культурно выделенными?, Как выделенные значения могут выражаться в данном естественном языке или в данном языке тела?.

4. Наконец, система отвечает на вопросы, связанные с характером, т. е. формой, смыслом, особенностями употребления, стилистическими свойствами и др. языковых и телесных знаковых выражений, которые отображают СО, их признаки и значения признаков. Поскольку значения многих признаков СО в разных языках выражаются прилагательными, в системе хранится большое число сочетаний существительных с синтаксически подчиненными им прилагательными, которые выражают значение признака или характеризуют сам признак. Впрочем, часто выражения, которые характеризуют признак, и выражения, которые передают значение признака, бывают синонимичными. Ср. *большой живот* и *большой размер живота*, *круглые глаза* и *глаза круглой формы* — хотя в сочетаниях *большой размер живота* и *глаза круглой формы* прилагательные *большой* и *круглый* синтаксически присоединяются к именам признаков, семантически они выражают значения этих признаков.

Завершая данный раздел, отметим свойство **открытости** обсуждаемой базы данных. Под открытостью имеется в виду возможность пополнения базы данными, которые позволят отвечать на новые содержательные вопросы, относящиеся к телу и телесности. Например, здесь мы ничего не говорили о типах гипотетических вопросов, которые касаются жестов разных семиотических классов и в реализации которых участвуют те или иные СО.

В перспективе предполагается дополнить нашу базу данных другими (по сравнению с уже имеющимися) видами информации. Например, мы хотим (1) добавить к имеющимся в базе текстовым иллюстрациям графические, схематические или рисуночные, в частности, ввести в базу изображения жестов, (2) снабдить базу системой разнообразных стилистических помет, (3) пополнить систему удобными для доступа модулями, в которых бы описывались тело отдельно мужчин и отдельно женщин, отдельно взрослых и отдельно детей и пр.

Разумеется, пока еще сложно сказать, какие в принципе нам могут понадобиться информационные блоки. Важно, однако, уже сегодня обеспечить ту степень гибкости и свободы, которые бы позволяли при необходимости пополнять систему без существенной перестройки ее внутренней структуры.

3. Компьютерная обработка запросов и возможности пользователя системы

Данная система создавалась для людей, которые интересуются ролью тела и СО в коммуникации, а главным образом для тех, кого интересует взаимодействие вербального и невербального семиотических кодов, в которых активную роль играет человек.

Наша база данных расположена в интернете на сайте http://c2897.ip4.ru. При обращении к системе с вопросами, или *запросами*, требуется знать особый

пароль и ввести в систему имя пользователя[12]. При входе пользователю открывается меню с пятью типами запросов: (1) «определение типа СО, к которому данный объект относится»; (2) «поиск всех СО, обладающих данным признаком»; (3) «поиск всех признаков, которыми обладает данный СО»; (4) «поиск всех значений для каждого из признаков, которыми обладает данный СО» и (5) «установление соответствия между языковым выражением, содержащим имя данного СО, и определенным значением некоторого признака этого СО».

Поясним более подробно, как обрабатываются системой некоторые из запросов.

СО группируются в содержательные классы по разным — структурным, физическим и функциональным — признакам. В частности, СО объединяются в типы. Отдельный тип СО образует тело человека. Есть типы «части тела», «части частей тела», «органы», «телесные покровы», «телесные жидкости», «кости», «железы» и др. (о типах, их особенностях и принципах их образования см. подробно в статье Крейдлин Переверзева 2011). Примерами вопросов о типах СО являются предложения *К какому типу соматических объектов относится палец руки?* и *К какому типу соматических объектов относятся ногти?*. Поскольку есть СО, которые могут относиться к разным типам (а следовательно, деление СО не является классификацией), ответов на вопрос о типе может быть несколько. Так, глаз — это и часть лица (т.е. часть части тела), и орган <зрения>, то есть глаз относится сразу к двум типам: первый характеризует глаз, так сказать, со структурной точки зрения, а второй — с функциональной. В построенной системе предусмотрены оба вида ответов на этот и подобные вопросы. Однако она может давать и дифференцированные ответы — в случае, если пользователь указывает в запросе вид признаков, по которым выделяется тип[13]. В данном случае, если запрос имеет примерно такой вид *К какому типу относится глаз со структурной точки зрения?*, то ответ будет *Часть части тела*, а если *К какому типу глаз относится с функциональной точки зрения?*, то ответ будет *Орган*.

Пользователь может задать некоторый признак СО, (например, признак «форма») и обратиться к системе с просьбой выдать ему все СО, имеющие данный признак (форму). В ответ на вопрос о форме система выдает множество СО, которые обладают этим признаком: {рука, нога, голова, лицо, нос, …}, но в это множество не попадут, например, объекты типа «телесные жидкости», потому что телесные жидкости формы не имеют.

Наконец, пользователь может обратиться к системе, задав ей некоторое языковое выражение, содержащее имя СО (в частности, войти в систему просто с одним-единственным словом, обозначающим СО). Грамматическая форма

---

[12] Ситуация сегодня такова, что на сайт могут пока что зайти только участники проекта, то есть, воспользоваться им может лишь очень небольшое число людей. Причина этого заключается не в том, что мы хотим сделать сайт закрытым. Наоборот, мы стремимся к максимальной его открытости; просто построение системы еще не закончено, и нам не хочется показывать пользователю полуфабрикат.

[13] Информация о видах признаков и об их отнесении к структурным, физическим или функциональным хранится в памяти системы.

выражения, подаваемого на вход, существенна, поскольку, к сожалению, пока еще в систему не встроены блоки синтаксического и морфологического анализа. Поэтому система выдает только наиболее частотные способы выражения значения признака СО и при этом накладывает существенные языковые ограничения на вид запроса. Так, имя СО, если он непарный[14], должно быть в форме единственного числа именительного падежа (а в случае, если СО — парный, такой как глаза, уши, ноздри и под., — в форме множественного числа), глагол должен быть в форме инфинитива несовершенного вида и под.

При реакции на запрос об установлении соответствия между языковым выражением, содержащим имя данного СО, и определенным значением некоторого признака этого объекта система ищет в памяти нужное языковое выражение, но находит его только в случае, если оно отражает значение признака данного объекта. Тогда система ставит в соответствие вводимому языковому выражению множество значений признака данного объекта (в предельном случае — одноэлементное). В случае, если системе задается просто имя СО, она ставит ему в соответствие множество всех значений всех признаков СО — референта данного имени.

Приведем примеры подобных запросов системе и возможных ответов на них. Предположим, что некто обратился к системе с выражением *сухая рука*. Поскольку в памяти системы имеется признак «каритивность <руки>»[15], одно из значений которого является /сухой/, система установит соответствие между выражением *сухая рука* и этим значением признака.

4. Общая архитектура системы и ее внутреннее устройство

Поскольку обычный пользователь не обязан знать, как система устроена, данный раздел статьи мы адресуем, прежде всего, тем людям, которым интересны технология построения и программирование баз данных.

С формальной точки зрения, наша система представляет собой совокупность связанных друг с другом таблиц (их количество на сегодняшний день более 25). Каждая из таблиц имеет свою структуру и наполнение. Чтобы описать внутреннее устройство базы данных, следует, таким образом, охарактеризовать все таблицы по отдельности и эксплицитно указать связи между ними.

Чтобы не усложнять техническими деталями описание системы в целом, мы опишем здесь лишь основные таблицы, которые демонстрируют основные

---

[14] О парных и непарных СО см. Крейдлин, Переверзева 2010.

[15] Признак «каритивность» (недостаточность), относящийся к некоторому СО, в том числе к телесному, означает полное отсутствие или частичная нехватка в данном СО какой-то существенной части. Под существенной частью имеется в виду та часть, без которой либо данный СО, либо обладатель СО в некотором отношении представляются неполноценными, несостоятельными (например, без этой части СО плохо функционирует, имеет аномальные (недостаточные) размер или форму или что-то происходит с его внутренними частями и т. п.). Подробнее о признаке «каритивность» см. работы Крейдлин, Переверзева 2010 и Толстая 2008. В первой работе вводится также признак «избыточность СО», противоположный признаку «каритивность СО».

элементы архитектуры системы, и раскроем некоторые важные связи между таблицами.

4.1. Внутреннее устройство таблиц

Среди основных таблиц есть та, в которой хранятся имена СО, причем не все возможные, а только стандартные, типичные. Другими словами, в этой таблице есть, например, слово *голова* (часть тела), но нет его словообразовательных производных, то есть слов *головушка*, *головка* и др.

Есть таблица, в которой хранятся имена признаков СО. Или имеются таблицы с информацией о наиболее частотных значениях каждого из признаков и о выделенных значениях тех же признаков. Среди других таблиц системы отметим ту, в которой указаны возможные признаки обладателей СО.

Входом в каждую таблицу является кортеж из нескольких элементов. Обязательными элементами каждого такого кортежа являются имя таблицы и идентификатор. Формально говоря, идентификатор — это множество, состоящее из трехэлементных кортежей (троек). В тройку входит порядковый номер элемента таблицы (это первый элемент кортежа), имя некоторой языковой или жестовой единицы, которая содержательно характеризует данную тройку (второй элемент кортежа), и название таблицы (третий элемент кортежа).

Пусть таблица имеет заголовок «имя СО». Ее идентификатором является множество троек следующего вида: <1, *голова*, имя СО>, <2, *рука*, имя СО>, <3, *нога*, имя СО> ... [16]

Связи между таблицами также эксплицитно фиксируются в системе. Например, таблица с номером 1, которая называется «имя СО», связана с таблицей с номером 2, — «признаки СО», и связь этих таблиц явным образом отмечается при каждой из них.

4.2. Интерфейс

То, о чем мы говорили в предыдущем пункте, касается исключительно разработчиков системы. Обычный же пользователь обращается к **интерфейсу**, т. е. к совокупности технических средств, которыми обладает система для обеспечения взаимодействия пользователей с различными ее устройствами и программами.

Чтобы диалог человека с системой протекал более комфортно, нами были составлены инструкции с указанием того, что должен делать пользователь при появлении на экране определенных изображений. Эти изображения в нашей системе имеют вид особых бланков, обладающих в общем случае иной структурой, чем таблицы внутри системы. Эти бланки и проще, и нагляднее.

---

[16] Поскольку элемент тройки «имя СО» регулярно повторяется можно “сократить” идентификатор до пары <1, *голова*>, <2, *рука*>, <3, *нога*>...

Каждая инструкция, в свою очередь, имеет вид последовательности команд, которые должен выполнить человек, желающий получить ответ на свой вопрос. Результаты поиска тоже помещаются в бланках другого рода (в **бланках запроса**), каждый из которых состоит из двух столбцов.

Левый, или первый, столбец у каждого бланка одинаковый и представляет собой меню запросов (меню включает в себя все те группы запросов, о которых мы говорили ранее). Правый, или второй, столбец у каждого бланка запроса в общем случае свой: его структура и наполнение зависят от выбранного типа запроса. В этом столбце пользователь формулирует свой запрос, после чего ждет от системы ответа.

Вот как примерно выглядит бланк запроса. Допустим, что человеку нужно найти некоторый соматический объект по определенному набору признаков. Он заходит в систему, и перед ним открывается страница с бланком запроса следующего вида[17]:

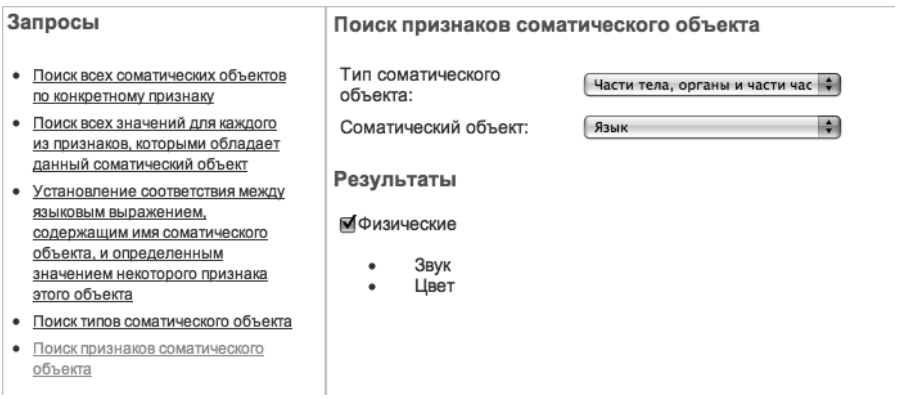


Рис. 1. Бланк запроса

Как показывает правый столбец бланка, пользователь остановился на запросе «поиск всех признаков, которым обладает СО язык». Продемонстрируем кратко, что он увидит в пункте «Результаты» (столбец с таким пунктом имеет каждый бланк запроса).

Информация, которую получит пользователь (см. фрагмент бланка на рис. 2), читается следующим образом: СО язык обладает признаками «звук <языка>» (действительно, *язык щелкает*, *цокает*), «ориентация <языка>» (ср. *показать язык 1* (имя жеста: кончик языка ориентирован на адресата), *показать язык 2* (свободное сочетание: например, человек показывает язык в кабинете врача; язык ориентирован на адресата), «размер <языка>» (*длинный*

[17] Система построена таким образом, что по умолчанию пользователь получает сразу бланк с запросом «поиск СО по данному признаку». В случае, если ему нужен не этот тип запроса, а какой-то другой, он его находит в меню, и правая столбец бланка запроса соответственно меняется.

язык), «каритивность <языка>» (*сухой язык*), «избыточность <языка>» (*влажный язык*), «текстура <языка>» (*шершавый язык*), «толщина <языка>» (ср. сочетание *толстый язык* в предложении *На стене ванной показывала толстый язык нарисованная смешная рожа* (*А. Лазарчук. Все, способные держать оружие…*)), «цвет <языка>» (*малиновый язык*; *белый язык*).

Результаты

Обладатель	Соматический объект	Значение признака
Физическое состояние: болезнь	Язык	Цвет: Белый

Рис. 2. Результаты

5. Заключение

В работе была представлена компьютерная вопросно-ответная система (база данных), содержащая информацию о человеческом теле и различных явлениях телесности. Было определено ее назначение, охарактеризовано ее внутреннее устройство и показаны ее информационные возможности. Эта система является моделью семиотической концептуализации тела, причем моделью, относящейся не только к русскому языку и русской культуре, но и к другим языкам и культурам.

Принципиальными достоинствами системы являются, на наш взгляд, ее открытость и связанная с нею возможность ее коррекции и пополнения. В то же время она обладает по меньшей мере одним очевидным недостатком, а именно отсутствием встроенного грамматического анализатора, что резко ограничивает языковую свободу пользователя при формулировке запроса к системе. Устранить этот недостаток планируется со временем после решения более актуальных на данный момент для системы задач. Также стоит отметить, что со временем система сможет учитывать частоту вхождений способа выражения признака, что позволит ранжировать результат.

Литература

1. *Кадыкова А. Г., Крейдлин Г. Е.* Части тела в русском языке и в русской культуре: признак «цвет». // Вестник РГГУ (Московский лингвистический журнал). — М.: Издательский центр РГГУ, 2010. — № 9/52 (т.12). — С. 47–64.
2. *Крейдлин Г. Е.* Невербальная семиотика: Язык тела и естественный язык. — М.: Новое литературное обозрение, 2002. — С. 190.
3. *Крейдлин Г. Е.* Лексикография жестов и их номинаций (словари и базы данных) // Материалы VII Международной школы-семинара «Современная

лексикография: глобальные проблемы и национальные решения». Иваново, 2007. С. 17–19.

4. *Крейдлин Г. Е., Летучий А. Б.* Части тела в русском языке и в невербальных семиотических кодах. II. Щеки. // Русский язык в научном освещении. — 2010. — № 1. — С. 222–235
5. *Крейдлин Г. Е., Переверзева С. И.* Ориентация тела и его частей: коммуникативные ситуации и стратегии поведения // Архипов А. В., Захаров Л. М., Кибрик А. А. и др. (ред.). Фонетика и нефонетика. К 70-летию Сандро В. Кодзасова. — М., 2008. — С. 589–602.
6. *Крейдлин Г. Е., Переверзева С. И.* Ориентация частей тела в русских позах и жестах: нормы и стереотипы // Стереотипы в языке, коммуникации и культуре. — М., 2009. — С. 468–481.
7. *Крейдлин Г. Е., Переверзева С. И.* Семиотическая концептуализация тела и его частей: тело, части тела и телесность // Теоретические и прикладные аспекты современной филологии: материалы XV Всероссийских филологических чтений имени проф. Р. Т. Гриб (1928–1995). Красноярск, 2010.
8. *Крейдлин Г. Е., Переверзева С. И.* Семиотическая концептуализация тела и его частей. I. Классификационные и структурные характеристики соматических объектов. Вопросы филологии. 2010. — №2 (35). — С. 42–51.
9. *Крейдлин Г. Е., Переверзева С. И.* Дружеские и любовные жесты. II. Поцелуи // Лингвистика для всех. Летние лингвистические школы 2009, 2010 и 2011. М.: МЦНМО, 2012 (в печати).
10. *Крейдлин Г. Е., Переверзева С. И.* Дружеские и любовные жесты // Лингвистика для всех. Летние лингвистические школы 2009, 2010 и 2011. М.: МЦНМО, 2012 (в печати).
11. *Летучий А. Б.* Часть тела/форма «кулак»: функции, концептуализация, место в системе частей тела // Вестник РГГУ (Московский лингвистический журнал). — М.: Издательский центр РГГУ, 2008. — № 6 (т.10). — С. 91–108.
12. *Толстая С. М.* Пространство слова. Лексическая семантика в общеславянской перспективе. — М.: Индрик, 2008.

SOME DIFFERENCES BETWEEN CZECH AND RUSSIAN: A PARALLEL CORPUS STUDY

Klyueva N. M. (netsie@yandex.ru)

Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic

We present a comparative study of some constructions in Czech and Russian. Though Czech and Russian are closely related Slavic languages, they have a few differences at the level of syntax, morphology and their semantics. We discuss incongruities that we found in a parallel Czech-Russian corpus, mainly reflecting differences in the sentence structure. The linguistic evidence presented in the paper will be used while constructing the transfer module of a rule-based machine translation system between Czech and Russian.

Key words: Czech, Russian, parallel corpus, language differences

1. Introduction

In this paper we make a corpus-based research in order to find the most frequent differences between Czech and Russian with the respect to the sentence structure. The languages share a lot of common features on every language level, still they have differences that might present a challenge for example to the Machine Translation system as well as to the language teaching.

The contrastive study of Slavic languages and our concrete pair of languages is a well studied topic. Just to name some of them, the paper [6] presents structure similarity measuring mainly on the material of Slavic languages. The authors in [2] compare Czech and Russian aspect and negation.

Here we neither suggest some global algorithm for finding differences nor focus in detail on one concrete linguistic problem, but we rather list the most frequent incorrespondences between Czech and Russian.

Our work was initially motivated by the task of creating transfer rules for the Machine Translation system between Czech and Russian [4] that can capture the main differences between the languages, and the results presented here are the first step towards the set of such rules.

The paper is divided into two main parts — in the Section 2 we provide a short description of a parallel corpus and a method how we searched for the examples. Section 3 lists the most frequent differences between the languages.

2. Parallel Corpus Study

One of the most popular technique of linguistic investigation now is the study of some language fact on an annotated corpus. We have made our research on a parallel segmented and tokenized Czech-Russian corpus[1] that contains about 100.000 sentences on each side. For our task we have chosen 88.000 sentences with a sentence alignment one-to-one, where one Czech sentence is aligned to one Russian.

The corpus contains news texts mainly with political, social or economic thematics downloaded from the site www.project.syndycate.org.

The corpus is tagged with a morphological tagger on both sides. For Russian we have used the TreeTagger [10] and for Czech language the Positional Tagger [3], the result annotation looks like follows:

(1cz) Chápu|chápat|VB-S---1P-AA--- jejich|jeho|PSXXXXP3-------
postoj|postoj|NNIS4-----A----
(1ru) Я|я|P-1-sn понимаю|понимать|Vmip1s-a-p их|они|P-3-pa
позицию|позиция|Ncfsan
'I understand their position'

As the two taggers exploit different annotation schemes, we did not make a full table of tag correspondences for Czech and Russian. We took only the first letter from the tag which reflects the word class and unified it for Czech and Russian[2] according the schema in [3], so we will not go into the detail of both annotation schemes as the rest information will not be used for this research . We assigned each sentence with a clue encoding order of sentence constituents (first letters of part of speech), ex (2).

(2cz)	*Chápu*	*jejich*	*postoj*	VPN:	(VerbPronounNoun)
	Understand.1Sg.Pres	their	position		
(2ru)	*Я понимаю*	*их*	*позицию*	PVPN:	(PronVerbPronNoun)
	I understand	their	position		

'I understand their position'

Therefore we have calculated a Levenshtein's[3] [8] distance between those sequences of part of speech tags that measures how different two strings (in our case the order of the word classes) are. Levenshtein's distance reflects the minimal edit distance — a number that shows how many edit steps need to be introduced into the Czech string to transform it to the Russian. So the sentences in the example (1) have the edit distance 1, which illustrates that to transform a Czech sequence of tags into the Russian only one letter (P-pronoun) should be added (which is actually the

[1] https://ufal.mff.cuni.cz/umc/cer/

[2] Abbreviations for POS in [3] are: N=noun, V=verb, A=adjective,P=pronoun, R=preposition, D=adverb etc.

[3] Levenshtein distance was calculated using Perl module https://metacpan.org/module/Text::Levenshtein

personal pronoun “Я” — I). So the more is the Levenshtein’s distance, the more are two sentences different from each other. Those sentences are therefore the main focus of our research. The statistics on the distribution of sentences with the respect to the distance is presented in Table 1.

Table 1. Distance between Czech and Russian sentences

Levenshtein's distance	**0**	**1**	**2**	**3**	**>3**
# of sentences	296	721	1503	2423	74372

It appeared that only a small amount of sentences have the same structure in Czech and Russian (edit distance 0), ex:

(3cz) *Oficiálně Čína zůstává komunistickou zemí.*
AdverbNounVerbAdjNoun
(3ru) *Официально Китай остается коммунистической страной.*
AdverbNounVerbAdjNoun
both: 'Officially China remains Communist country’

This fact was really astonishing because we expected much more correlations in the sentence structure.

One of the reasons might be that Czech and Russian sentences in this corpus are translated from English original in different, often novel way, ex.(4).

(4cz) *Evropa je krátkozraká* NounVerbAdj
Europe is myopic
(4ru) *Европа ведет себя недальновидно* NounVerbPronAdj
Europe behaves itself without foresight
'Europe is being myopic'

Even it is not always the fact that sentences with the distance is 0 will have similar structure. It is often the difference in phraseology that becomes the source of the incorrespondence. In the example (5) Czech and Russian use totally different constructions — fixed phrase units 'v sázce je' vs. 'на карту поставлена', though the sequences of parts of speech are the same:

(5cz) *V sázce je bezpečnost lidstva* PrepNounVerbNounNoun
At stake is security mankind.gen
(5ru) *На карту поставлена безопасность человечества* PrepNounVerbNounNoun
On map put.passive security mankind.gen
'At stake is the security of mankind'

Moreover, the news sentences are generally long and have complicated structure which will increase the amount of incorrespondences.

The third factor is that we have studied only the order of part of speech sequences. Probably if we have a more deep annotation the percentage of sentences with corresponding structure will be higher. More deep syntactic analysis would have help to detect more non-trivial differences and to handle such incorrespondences as for example ellipsis. To provide such an analysis, the high-quality syntactically annotated parallel treebank would be needed, which we do not have for both Czech and Russian. From the experience of the Prague Czech-English Dependency Treebank[1], a lot of manual annotation work should be used. Moreover, the parallel texts chosen for the treebank should be translated manually with the instructions for annotators to translate as close as possible to the original, so that it would be easier to align the annotated trees. For this goal it is not sufficient to take parallel texts already translated as the sentences can be translated from one language to another in a novel way.

As we do not have so far human resources for such a long-lasting goal, we made our study on what we have — a parallel corpus with simple morphological annotation — exploiting the corpus only in a linear word-for-word manner.

2.1. Differences in a Sentence Structure between Czech and Russian

To illustrate the cases where Czech and Russian use the different construction we have taken those sentences that have the Levenshtein's distance 1, 2 or 3. They reflect some of the relevant differences in the sentence structure and at the same time do not overload the sentence with too much incorrespondences. Below we describe the most frequent differences according to the parallel corpus. If there exist several options to translate a construction from Czech into Russian, they are given with an appropriate statistics from the corpus.

2.2. Structural differences

One of the most big challenges while translating from Czech into Russian is the omission of a **subject pronoun** in Czech and a verb 'to be' which is dropped in Russian (6ru) but is present in Czech[4]. According to the corpus, in 1102 cases the copula construction in Czech corresponds to a zero-copula construction with a dash symbol in Russian, in 1673 cases there is no dash and no verb like in (6ru). The copula verb was translated as a verb 'являться' — 'appear to be' in 1331 cases(7ru, first clause). This verb occurs generally in the written texts and sounds officially.

(6cz)	*Vlády*	*jsou*	*zkorumpované*	NounVerbAdj
	Goverments	are	corrupt	

[4] In the deep syntactic analysis within the treebanks the missing sentence elements are generally annotated with zero mark Ø .

(6ru) *Правительства Ø коррумпированы* NounAdj
Governments corrupt
'Governments are corrupt'

(7cz) *První strategie je krátkozraká a druhá je ošklivá .*
First strategy is shortsighted and second Ø nasty.
(7ru) *Первая стратегия является недальновидной , а вторая — отвратительной .*
First strategy is shortsighted and second — nasty.

Analytical past in Czech is formed by the appropriate form of the verb "to be" and the past participle whereas in Russian the copula is omitted as in the example (8).

(8cz) Přišla jsem pozdě *VerbVerb_auxAdv*
come.Past.Fem to_be.1Sg.aux late
(8ru) я пришла поздно *PronVerbAdv*
I.1Sg come.Past.Fem late
'I came late'

***Reflexive particle** in Russian is incorporated into a verb, and in Czech — though considered to be a part of a lemma [5] — is written separately from the verb:*

(9cz) *Proč* ***se*** *Shiller mýlil?* PronPartNounVerb
Why refl.part Shiller mistake.3Sg.Past
(9ru) *Почему Шиллер ошиб**ся**?* PronNounVerb
Why Shiller mistake.3Sg.Past
'Why was Shiller wrong?'

*It is sometimes misleading when the particle stands far away from the verb and it can not be easily identified to which verb it belongs or even in case if the particle '**se**' if it is really a particle or a vocalized preposition 'se' (with):*

(10cz) Popsali nové odrůdy rýže s, o něž jsou připraveni **se** se svými africkými protějšky **podělit** .
Described new sorts of rice, about which are.3Pl ready se.refl se.prep their African colleagues share.inf

(10ru) Они рассказали о новых сортах риса , которыми они готовы **поделиться** со своими африканскими коллегами .
'They described new sorts of rice, with which they are ready to share with their African colleagues'

Often things get complicated when some of the listed phenomena (ex. pro-drop, past tense, reflexive verbs) got gathered in one sentence and the clitics go in Wackernagel's position which mixes the sentence structure even more:

(11cz)	Dlouho	jsem	se	smál.	*AdvAuxReflVerb*
	Long	*to_be.aux*	*refl.particle*	*laugh.Past*	
(11ru)	*Я долго*			*смеялся.*	*PronAdvVerb*
	I long			laugh.Past	
	'I laughted for a long time'				

It would have been more appropriate to make a deep syntactic analysis of the evidence in the above examples (6)–(11) within either Prague Dependency Treebank[5] or SynTagRus treebank of Russian[9], where those 'trivial' differences would be eliminated on the deepest level of annotation. However, we should have a big amount of annotated parallel data, which is difficult to built and it is not realistic so far.

On the clause level the obvious difference is the usage of some coordinating **conjunctions with contrastive meaning,** namely the order of elements in such clauses.

(12cz)	*Trest*	*však*	*mohl*	*být*	*tvrdý*	NounConjVerbVerbAdj
	Punishment	but	might	be	hard	
(12ru)	*Но наказание*		*могло*	*быть*	*суровым*	ConjNounVerbVerbAdj
	But punishment		might	be	hard	
	'But the punishment might be hard'					

(13cz)	*Nejprve*	*ale*	*byl*	*chaos*
	First	but	was	chaos
(13ru)	*Сначала ,*	*однако ,*	*был*	*хаос*
	First ,	but ,	was	chaos
	'First it was a chaos though'			

The Czech contrastive conjunction *však* usually takes the second position in the sentence (12cz), which causes dissimilarity with the Russian translation equivalent(12ru) . The conjunction *ale* may also take the second position (13cz), so the order of elements may be similar in Russian (13ru). There is also a special meaning of the word *ale* that expresses amazement or surprise (14) that does not exist in Russian:

(14cz)	*To*	*byla*	***ale***	*cesta!*
	That	was	but	trip
(14ru)	*Ну*	*и*	*дорога*	*была!*
	Well	and	trip	was
	'What a trip was it!'			

According to the corpus statistics Russian tends to use a sentence with the interrogative particle '**ли**' more often than Czech with the respective **-li**. This particle occurs in 1873 sentences in Russian and only in 208 for Czech both in interrogative sentences and relative clauses. Instead, in Czech sentences other particles with similar meaning are used: zda — 454 translations, ex.(15), jestli — 37, or there is no particle at all in 993 cases, ex. (16).

(15cz) O*tázka tedy nezní , **zda** Evropa existuje , ale **zda** jsme spokojeni s tím , jak funguje .*
(15ru) *Вопрос заключается не в том , существует **ли** Европа , а в том , удовлетворены **ли** мы тем , как она функционирует .*
'The question is not whether Europe exists, but if we are content with the way it function '

(16cz)	*Praskne*	*další*	*bublina?*	VerbAdjNoun
	Burst	next	bubble?	
(16ru)	*Лопнет ли*	*очередной*	*пузырь?*	VerbParticleAdjNoun
	Burst if(int. part.)	next	bubble	

'Will the next bubble burst?'

2.3. Differences in connection with lexicology and idioms

A multilingual expression or a certain fixed lexical phrase in one language is translated into the other language in the other way or just can be translated descriptively. Here we suggest three examples of differences that belong to the field of phraseology and idiomatics.The differences in lexicology are not easy to detect automatically, so here we provide several created examples not from the corpus.

The sentence (17cz) reflects that the woman has not done the action herself, but someone else did it. In Russian (17ru) it can not be identified from a sentence whether the hair was cut by woman herself or by someone else, but native speakers know, that generally it is the second option. The Czech causative construction (17cz) can sound strange to the native speaker of Russian, and vice versa a Russian variant may seem awkward and even funny for the Czech speaker (a lady cutting her hair herself at a hairdresser).

(17cz)	Nechala	si	ostříhat	vlasy	*v*	*kadeřnictví*
	Let.Past.Fem.Sg	*refl.part*	*cut*	*hair*	*at*	*hairdresser*
(17ru)	*Она*		*подстригла*	*волосы*	в	парикмахерской
	She		cut.Past.Fem.Sg	hair		

'She had her hair cut at hairdresser'

In (18) it is not so obvious which Russian equivalent should be used to the Czech construction '*slyšet na*' — 'hear at' as this meaning of the verb 'hear' is specific to the Czech language only.

(18cz)	Rusové	slyší	na	české lázně	
	Russians	*hear*	*at*	*Czech*	*spa*
(18ru)	Русские	интересуются		чешскими	курортами
	Russians	interest.3Pl		Czech	spa

'Russians like Czech spa' or 'Russians are interested in Czech spa'

Idiomatic expressions present the big challenge for our task because their word-for-word translation will result in a construction that may sound awkward or even ungrammatical in another language:

(19cz)	*agenda*	*svobody*	*nezmění*	*přístup*	*lidí*	***přes***	***noc***
	agenda	freedom.gen	change.neg.fut	attitude	people.gen	**over**	**night**
(19ru)	*план*	*свободы*	*не изменит*	*отношение*	*людей*	***в один миг***	
	agenda	freedom.gen	not change.fut	attitude	people.gen	**in one second**	

'a freedom agenda will not change people ' s attitudes overnight'

3. Conclusion

In this paper we have presented a comparative study of Czech and Russian with the respect to the differences mainly in the sentence constructions. The examples illustrating differences are based on the parallel corpus research. The method of finding differences proposed here was not sufficient enough as it showed a very low percentage of correspondences in sentence structure which intuitively should be much higher. Still, we have selected some frequent differences in Czech and Russian that may serve as a basis for the detailed research in the future involving more deep syntactic analysis .

The evidence described here is also planned be exploited in the Rule-Based Machine Translation from Czech into Russian.

References

1. *Curin J.,* et al. (2004) Prague Czech-English Dependency Treebank 1.0 Linguistic Data Consortium, Philadelphia, 2004
2. *Dickey S. M. and Kresin S. C.* (2009) Verbal aspect and negation in Russian and Czech. Russian Linguistics 33(2):121–176.
3. *Hajič J.* (2004). Disambiguation of Rich Inflection (Computational Morphology of Czech).Karolinum, Charles University Press, Prague, Czech Republic, 2004.
4. *Hajič J., Hric J. and Kuboň V.* (2000) "Machine Translation of Very Close Languages". Proceedings of the sixth conference on Applied natural language processing. Seattle, Washington pp. 7–12
5. *Hajič J.* et al. (2006) Prague Dependency Treebank 2.0 Linguistic Data Consortium, Philadelphia, 2006 .
6. *Homola P. and Kubon V.* (2006) A Structural Similarity Measure. Proceedings of the Workshop on Linguistic Distances, pages 91–99, Sydney.
7. *Lopatková M., Žabokrtský Z. and Benešová V.* (2006) Valency Lexicon of Czech verbs VALLEX 2.0. UFAL Technical Report TR-2006-34, 27+233 p., 2006
8. *Levenshtein V. I.* (1966) Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics. Doklady

9. *Nivre J., Boguslavsky I., and Iomdin L.* (2008) Parsing the SynTagRus treebank of Russian. In Proceedings of the 22nd International Conference on Computational Linguistics — Volume 1 (COLING '08), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 641–648.
10. *Sharoff S.* "Russian tagset and Russian statistical taggers", available at: http://corpus.leeds.ac.uk/mocky/

HUMAN CHARACTERS THROUGH THE PRISM OF ADVERBS

Kobozeva I. M. (kobozeva@list.ru),
Lukashevich N. Iu. (natalukashevich@mail.ru)

Lomonosov Moscow State University
Department of Psychology
Department of Philology, Moscow, Russia

The present paper deals with adverbs describing actions of a person from the character trait perspective — words like *carelessly*, *tactfully*, *bravely*, etc, derived from adjectives predicating to a person some character trait (*careless*, *tactful*, *brave*). The aim was to see what information such adverbs provide for defining patterns of behaviour that constitute the meaning of corresponding character trait adjectives. It is argued that the analysis of adverbs' contexts helps to outline more precisely the range of situations relevant for manifestation of a certain character trait, the range of actions which a person with the described character trait is inclined to perform, and their motivations. This is demonstrated by the analysis of contexts with *sincerely*, *frankly* and *candidly* from the British National Corpus.

Key words: cognitive semantics, agent-oriented adverbs, behaviour pattern, character nominations, synonyms

The present paper deals with adverbs like *carelessly*, *tactfully*, *bravely*, etc here referred to as character adverbs (ChAdvs) because they describe actions of a person as displaying a certain character trait. They are mostly derived from character adjectives (ChAdjs) that predicate the corresponding character trait to a person.

An important feature of this group of adjectives is that people attribute a character trait denoted by them to a person by generalizing from specific cases of the person's behaviour. When an adverb is used, a specific action of a person is qualified as belonging to such a type of cases, which forms the basis for such generalizations.

From the lexicographic point of view both adjectives and adverbs from this semantic field present a certain problem. Attempts to provide an adequate representation of the meaning of such words based only on traditional methods of semantic analysis often prove to be unsuccessful.

For instance, dictionaries frequently introduce meanings of ChAdjs by providing a list of synonyms or by making a reference to the meaning of another member of the same synonymic row. Such an approach to meaning representation allows to emphasize similarities between synonyms, leaving the differences unclear. One will find an example of this in the definitions provided in Longman Dictionary of Contemporary English (1995) for such adjectives as *sincere, frank, candid, honest, open, and truthful*:

Sincere	someone who is sincere is honest and says what they really feel and think
Frank	honest and truthful in what you say
Candid	directly truthful even when the truth may be unpleasant or embarrassing
Open	honest and not wanting to hide any facts from the people
Honest	someone who is honest does not lie or steal
Truthful	someone who is truthful does not usually tell lies

Adverbs are treated in a similar way, though some of them are not represented as separate lexical entries:

Sincerely	in a sincere way
Frankly	honestly and directly, especially in speech; [sentence adverb] used to show that you are saying something direct and honest
Candidly	–
Openly	in a way that does not hide your feelings or opinions
Honestly	1) in an honest way ("I don't know", she answered honestly.) 2) used to say that you really think that something is true, especially when it seems surprising (Does he honestly expect me to believe his story?)
Truthfully	–

For adverbs the situation is even more complicated, as they are generally considered to have a meaning which can be produced by a standard derivation operation from the meaning of the producing adjective. (Because of that some of them are not represented separately, as it has been shown above, but rather mentioned as a derivative in the entry for the corresponding adjective.) However, as it has been noted ((Penkovskii 1988), (Filipenko 2003)), this is not always true, and the divergence of meaning often makes it necessary to have a separate lexical entry for an adverb in the dictionary even in cases when it follows regular word-formation patterns.

The fact that adequate representation of meanings is problematic for words belonging to this semantic field can be explained by certain features of their semantics, such as a significant subjective part shaped by personal experience, an evaluative component in their meaning, etc (Lukashevich, Kobozeva 2011). Another important point is that to understand and use such words correctly people need to be able to compare and relate an individual instance of behaviour with an existing generalized scheme. Therefore an adequate representation of such word's meaning should somehow reflect this.

In psychology a trait of character is often regarded as a behavioural stereotype which is realized with high probability in a situation relevant for exhibiting this trait. Therefore setting a relevant situation and a stereotype of behaviour in such conditions seems the most appropriate way to represent the meaning of a word denoting a character trait (Lukashevich 2004). Such a representation tool named "behaviour pattern" («shablon povedeniia») was introduced in (Martem'ianov, Dorofeev 1969), (Martem'anov 1999) for the purposes of automatic language processing. A behaviour

pattern, associated with a certain character trait is a generalized implicative scheme establishing a link between the initial relevant situation and the stereotyped behavioural response of a person with this trait of character. Providing such behaviour patterns with prototypical ('best') examples of real-life situations and specific behaviour in them would allow to account for the ability of a native speaker to make inferences quickly in real life (Lukashevich 2002, Lukashevich 2004).

To obtain information about typical and real-life situations thorough and extensive analysis of the ways in which such words are used needs to be carried out. To be informative from this point of view, the text should contain a more or less clear description of the actions, or types of action performed by a person with a particular character trait, and of conditions which trigger them. Such details are not often explicitly present in the text where a ChAdj characterizes a person. Because of that analyzing contexts which illustrate the use of ChAdjs proves very time- and labour-consuming. It may even come to a standstill at a certain point, when synonymic characteristics are being studied. In this case analyzing contexts with adverbs, a class of words primarily used as predicate modifiers, may provide a look from a different perspective and possibly a way out of such an impasse. It is especially important that ChAdvs are used with verbs directly naming human actions and states that result in assigning the corresponding character trait to their subject.

The aim of this research was to have a closer look at ChAdvs and see what information these adverbs could provide for defining the pattern of behaviour and its best examples.

Many ChAdvs have a variety of uses. Their syntactic-semantic potential is illustrated by *frankly* in (1):

(1) (a) **Frankly**, I don't care.
(b) Many participants **frankly** admitted that they had very low expectations at the outset of the course.
(c) May I speak **frankly**?
(d) The result now looks **frankly** old-fashioned.

(1a) exemplifies a parenthetical use of ChAdv as a sentence adverb. It should be noted here that many adverbs (e. g. *luckily*, *interestingly*, *unfortunately*) may be used parenthetically in their sentential use. In this use ChAdvs belong to the class of speech-act speaker-oriented adverbs (*frankly*, *honestly*, *candidly*, *simply*, *seriously*) ((Jackendoff 1972), (Bellert 1977), (Ernst 2002)). They are singled out on the basis of the fact that they express the attitude of the speaker and / or a metatextual comment to what is being said and thus function as discourse markers in such contexts (e. g. in (1a) below the speaker emphasizes that he is saying what he is really thinking, that it may be unpleasant for the hearer, etc.). In this meaning they should be analyzed as modifying the act of speaking which is implicit in the semantic structure of any sentence. It needs to be pointed out that parenthetical usage of ChAdvs is treated in some dictionaries as a separate meaning in the adverb's lexical entry. Although such "speaker-oriented" uses are obviously semantically connected to other uses of ChAdvs, they will not be taken into account in this paper.

(1d) illustrates the use of an adverb as a modifier of a word denoting a property — an adjective or another adverb. In such cases the main meaning of ChAdv may undergo a major change called rebranding in (Karpova, Rakhilina, Reznikova, and Ryzhova 2011), thus here *frankly* becomes functionally close to a degree adverb[1], although its semantic ties with the main meaning 'in a way that is typical of a frank person' are not altogether broken. What is essential for us is that from (1d) one cannot deduce that *he* behaved as a frank person in the given case and that is why we ignore such uses of ChAdvs.

All of ChAdvs seem to have uses like (1b) and/or (1c), pertaining to their main meaning. In such cases they fall into the group of agent-oriented adverbs in accordance with the classification provided in (Ernst 2002): *cleverly, greedily, tactfully* etc.

Like most other adverbs, many agent-oriented adverbs have two interpretations, depending on whether the adverb modifies the predicate (as in (1c)) or the whole sentence (as in (1b)). These two interpretations referred to as "upper' and "lower" may be represented in the spirit of (Ernst 2002) by the paraphrases (1b') of (1b) and (1c') of (1c):

(1b') The *fact* that many participants admitted that they had very low expectations at the outset of the course *characterizes* them as frank.

(1c') May I speak in a *way* that will *demonstrate* that I am frank?

This scope difference was first mentioned in (Thomason, Stalnaker 1973) and later discussed in various other works (e. g., (Bellert 1977), (Ernst 2002)). It was thoroughly studied with regard to Russian adverbs in (Filipenko 2003), where a distinction was made between adverbs which can be used both as sentence modifiers and as predicate modifiers (as in (1b) and (1c)) and adverbs which do not show such flexibility and can only be predicate modifiers. The former group was named adverbials with floating scope, the latter — adverbials with fixed scope. Whether an adverb will allow various usage or not depends on its semantics. According to the analysis provided in (Filipenko, 2003) adverbials with floating scope characterize the situation "from the outside" (and do not characterize its semantic participants). That is why they can relate to various elements of the situation (the choice of which is governed by the communicative organization of the utterance) and can be used with verbs belonging to various semantic groups. As for adverbials with fixed scope, they characterize the situation as such, or represent one of its inner characteristics. They are linked with a specific semantic feature of the predicate, as they describe a participant (often optional) of the situation that the predicate denotes, and this participant is fixed for a particular adverbial. (For example, for *energetically* it is "the amount of energy" with which the action is performed.)

As for ChAdvs, they seem to be found in both of the above mentioned classes. This is definitely so for Russian, and the analyzed data for English prompt that this is the case for English as well. (Unlike *candidly* and *frankly*, contexts with *sincerely* seem to contain no evidence of sentential non-discourse marker use.).

[1] Surprisingly, this meaning is mostly not mentioned in the dictionaries, with a rare exception of Merriam Webster Online where it is represented as 'unmistakably evident' in *frank*.

No matter whether ChAdvs have upper or lower interpretation, both types link either the whole situation described in the sentence or some of its parameters with the character trait they refer to. Therefore, they can provide valuable information on the features of the relevant situation related to a character trait. Moreover, while focusing on different aspects of such situation (some of which may be optional), they have a potential of giving a fuller and a more detailed picture of the situations involved than the corresponding adjectives.

Thus for this research scope differences are not so important and do not set any limitations on the choice of contexts to be analyzed.

In order to see how the study of adverb usage can help in defining typical situations relevant for particular traits of character three adverbs belonging to the so called Candidness frame (FrameNet) — *sincerely*, *frankly* and *candidly* — were analyzed. Altogether 141, 431 and 67 contexts from British National Corpus (BNC) were examined respectively. Such disproportion in numbers can be accounted for by the following points. First of all, BNC contains only 67 sentences with *candidly*. To make the data comparable, more or less similar numbers of sentences containing *sincerely* and *frankly* were to be singled out. For *sincerely* all sentences containing "Sincerely Yours" and similar clichéd expressions used in letters were eliminated as irrelevant for the present research. Initial analysis of data for *frankly* revealed that up to 70% of sentences represent *frankly* and similar expressions (*quite frankly, frankly speaking*, etc) used as a discourse marker. Then in the remaining 30% about half of sentences demonstrate a different meaning of *frankly* used as in (1d) above. This meaning is also irrelevant for the purposes of the present research.

The verbs which were used in BNC contexts with *sincerely*, *frankly* and *candidly* can be roughly grouped as follows:

		sincerely	**frankly**	**candidly**
		91	89	61
I	**Communicative acts:**	**28**	**73**	**58**
1	'inform' (say, reply, confess, etc)	7	53	31
2	'describe smb or smth'(describe, write of smb etc)	–	1	6
3	'express an emotion / a wish' (express hope /dissatisfaction, apologize, thank, lament, ask, plead etc)	15	–	1
4	'express or demonstrate an evaluation':	3	2	2
4a	'positive' (praise, be appreciated)	2	–	–
4b	'negative' (blame, reproach oneself, look down on smb)	1	2	2
5	'accept and agree unwillingly that smth is true' (admit, recognize)	–	11	14
6	non-verbal communication (meet smb's gaze, smile, stare etc)	–	4	2
II	**Mental attitudes, states and activities:**	**58**	**14**	**1**

		sincerely	**frankly**	**candidly**
1	Beliefs and intellectual activities (believe, hold a viewpoint, mean, doubt; recognize 'realise', see smth as 'consider', judge smb's abilities be committed to etc)	17	4	1
2	Volitive and emotional states and attitudes (feel, like, respect, desire, hope, dread, be amazed / unhappy etc)	41	10	–
III	**Social actions that are not or not only communicative (vote, behave, open one's home 'exercise hospitality', hold one's glass out etc)**	**5**	**2**	**2**

The table above shows that verb distribution shows great similarity for *frankly* and *candidly*, whereas *sincerely* is rather distinctly different from them.

All the three characteristics are often displayed in communication activity, either verbal or non-verbal, but *frankly* and *candidly* are almost always linked to communication, while *sincerely* is much more often compatible with verbs denoting mental states and activities. It means that sincerity characterizes not only communicative behavior but also the way a person is committed to his beliefs, intentions and feelings.

Our three ChAdvs co-occur with verbs denoting representative speech acts (see line I.1 above), but the fact that *describing* smth does not co-occur with *sincerely* made us look closer at the cases with other representatives, *say, speak* and *tell*. It turned out that in all the cases with *sincerely* the verbs were used either for referring to speaking in general, as in (2a) or introduced utterances with illocutionary force of expressing evaluation and / or emotional state or attitude, as in (2b,c):

(2) (a) …they should speak sincerely, relevantly and clearly, while providing sufficient information…
(b)… 'That's marvellous,' she said sincerely…
(c) … 'He's the luckiest devil in the world,' he told her sincerely, taking her hands in his.

Thus we can conclude that *sincerity* is a character trait associated with a situation of having a subjective attitude towards smb or smth (e. g. *believe* vs. *know*) and a situation of expressing such an attitude, especially volitional, emotional or evaluative.

Our tentative definition of the character *sincere* is presented in (3):

(3) A person X is called *sincere* if the following is the case:

- whenever X has a belief, intention, wish, emotion or feeling, X is serious about it, that is X's state or attitude is genuine and conscious, not fake, imitated or shallow.
- whenever X expresses his belief, intention, wish, emotion or feeling, one can be sure that X really has it.

As for the evaluative aspect, *sincere* is a positive characteristic of a person or of a particular action described as demonstrating this quality. This can be explained by the fact that it is generally believed to be good not to lie about one's feelings.

With regard to *frankly* and *candidly* it can be said that they are both mostly realized in speech and both often used with verbs like *admit* meaning that a person either discloses or confirms unwillingly that something is true.

What makes *frankly* different from *candidly* and likens it to *sincerely* is its ability to co-occur with verbs denoting mental attitudes, states and activities including both intellectual activities (e. g. *see smth as smth* 'consider') and emotional states (e. g. *dread*, *be amazed*). It should be noted that in the latter case (exemplified in (4) below) although *frankly* is modifying a form of a verb and not an adjective as in (1d) the meaning in (4a) and (4b) is closer to (1d), than to (1b) or (1c),cf.:

(4) (a) Prince Philip was **frankly** disappointed in his first born and took no pains to hide his feelings.
(b) She was **frankly** dreading any further confrontation with the other girl.

In (4) *frankly* is used to characterize the emotional state of an experiencer as intense, not to classify it as typical of a frank person. This is not so with *sincerely*, as is can be seen from (5):

(5) We instantly lost all respect for the team and were *sincerely* disappointed.

Here the experiencer of the disappointment characterized himself and his companions as sincere, i. e. really experiencing the emotion and not just feigning it. And if we look at (6):

(6) I must confess I was … highly delighted when er you invited me to speak to you here today and er I'm **very** and **sincerely** pleased that you did.
we see *sincerely* conjoined with an intensifier *very*, which demonstrates that it means something different from mere intensification.

Thus the examples like (4) could be dismissed. Still it can be shown that there is a natural semantic link between a degree meaning of *frankly* and its main meaning, and this link is provided by one distinctive feature of a frank behaviour. From our analyses of data it can be concluded that among other things a frank person acts so as to make other people know something about himself or others that people normally hide. This is exemplified by (1b) where the agent can choose whether to say something or not and can also choose the appropriate way to do it. In (4) an uncontrolled negative emotional state is described. When such a state is intense it usually manifests itself (even if the experiencer would prefer to hide it) and so the inner state becomes known to the observer as if the experiencer informed him about it (cf. the meaning 'unmistakably evident' of *frankly* mentioned in footnote 1).

It is clear that for *frankly* and *candidly* the results above on their compatibility with semantic classes of verbs do not provide enough information for clarifying the

nature of the corresponding character traits (which mirrors the situation with adjective contexts from BNC).

A more detailed analysis of the actions described as *frankly* and *candidly* performed was carried out in order to find out in what ways the two corresponding characters differ from each other (if they do). We singled out common and distinctive features of the actions. As we expected from the start the majority of them turned out to be common for the two traits of character, among them:

1) the speaker says something that the addressee does not know and that is not obvious;
2) the information may be about the speaker (and / or his people), about the addressee (and / or his people), and about a third party;
3) the information disclosed may be a fact, an opinion, an evaluation, a wish, an intention, a plan, an emotional state, a feeling;
4) what the speaker says truly reflects what he thinks, wishes or feels;
5) the information in the overwhelming majority of cases contains negative evaluation (in the form of assertion, presupposition or entailment);
6) revealing the information may be against the interests of the speaker;
7) the speech act itself (the fact of saying what is said) in the overwhelming majority of cases negatively affects the image (known as "face" in politeness theory) of the speaker, the addressee or the third party;
8) the speaker may act as a private person or as an official representative of an organization;
9) the communication may be interpersonal or public.

Still definite preferences of usage can be observed: there is a tendency to prefer *frankly* in case when the negative information is about the speaker[2] and / or people or things he is personally identified with, as e. g. a government official with a government policy, see typical examples in (7):

(7) (a).… during the first interview Hepburn had denied any involvement, but when seen again 40 minutes later '**frankly** said that he was responsible for the assaults on both girls'.
(b) … It must be **frankly** recognized that there is at present no means of providing adequate protection for the people of this country against the consequences of an attack with nuclear weapons.

Incidentally, the only two examples of a speaker confessing to a highly intimate and commonly reprehensible (or at least not to be announced) thoughts and wishes contain *frankly* as a modifier.

On the contrary *candidly* prevails when the negative information is about others, as in (8):

[2] Negative information about X is an information that induces a negative evaluation of X.

(8) a. As you yourself so **candidly** pointed out, she was only marrying me for my money…
b. Renault's communications director Phil Horton **candidly** blames supply constraints because of strong French domestic demand, and, hitherto, a lack of diesel interest within Renault U. K.

Thus we can hypothesize that at least prototypically *frankness* is associated with revealing personal information, negatively affecting the speaker's face, while *candidness* is displayed mostly in speaking about the faults and drawbacks others.

One more distinction between the two ChAdvs is hinted at with the word 'direct' in the dictionary entry of *candid* cited above[3]. Indeed, the difference lies in the manner of speaking. Among the BNC examples of the *candidly* performed speech acts we find more expressive evaluative expressions (e. g. *half-fulfilled and half-frustrated, a third-rater; a manipulative little bastard; terrific* etc), direct expressions of evaluative or emotional attitude (e. g. *candidly dislike / express dissatisfaction /say 'j'aime Eva'* etc), categorical formulations (e. g. *Bob Dutton would never get a medal; everybody said that; Stephen's got more soccer skill than I ever had*, etc). For *frankly* the 'direct' (expressive, unequivocal, categorical) manner of speaking is irrelevant, as in (9):

(9) Her employer commented **frankly**, but **with great gentleness**, on his observations of Muriel's way of handling-others, particularly when conflict arose.

Evaluation connotations of *frank* and *candid* (in those instances where these words label human behaviour) may be different depending on the context. It logically follows from the fact that expressing your opinion or stating what you believe to be true may not always be good for all the participants of communication.

We do not formulate definitions of *frank* and *candid*, because evidently more data is needed and specially designed experiments with native speakers should be carried out. However, though the present research results in only a hypothesis, this can also be regarded as an achievement, because it gives the basis for further analysis. (For example, such was the situation with Russian characteristics *upornyi* and *nastoichivyi*, when a hypothesis formulated on the basis of verb distribution analysis was successfully used to make up and conduct an experiment which helped to identify in which cases native speakers would tend to choose one characteristic over the other (Lukashevich 2004).)

It would be relevant to note here that information about the evaluative component of ChAdjs' and ChAdvs' meaning may be of use in such spheres as sentiment analysis and opinion mining. With regard to the words discussed above, it can be concluded that *sincerely* as a rule indicates a positive assessment of a particular action. As for *candidly* and *frankly*, the evaluation they express is context-dependent and more ambiguous, but further research may reveal certain regularities there as well

[3] Strangely enough, this attribute, although absent in the explication of *frank*, appears in the explication of *frankly*, thus marring the difference between *frankness* and *candidness*.

(e. g. when describing actions related to talks and negotiations, they are always used as a positive characteristic).

It can be concluded that in general, compared to ChAdjs, ChAdvs provide a clearer picture of the range of behavioral patterns, associated with the character trait under analysis. Such information is of particular importance when there are several synonymic character trait nominations closely linked. Although sometimes a mere analysis of verb distribution is also not enough, and a deeper semantic analysis of contexts is required, adverbs are capable of providing valuable information for defining behaviour patterns and real-life situations instantiating them.

References

1. *Bellert I.* 1977 On semantic and distributional properties of sentential adverbs. Linguistic Inquiry, 1977, v.8, pp.337–350.
2. *Ernst T.* 2002 The syntax of adjuncts. Cambridge: Cambridge University Press, 2002.
3. *Filipenko M. V.* 2003. The semantics of adverbs and adverbial expressions [Semantika narechii I adverbial'nykh vyrazhenii]. Moscow, 2003.
4. *Jackendoff R.* 1972 Semantic interpretation in generative grammar. Cambridge, Massachusetts.
5. *Karpova O. S., Rakhilina E. V., Reznikova T., and Ryzhova D. A.* 2011. Meaning of estimation in semantic shifts of rebranding type in adjectives and adverbs (on the material of the database of semantic shifts in Russian adjectives and adverbs). Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoii Konferentsii "Dialog" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2011") (Bekasovo, 25–29 May 2011). Moscow, 2011. pp. 292–304.
6. *Lukashevich N. Iu.* 2002. Predicates describing character traits and behaviour patterns [Kharakterologicheskie predikaty i shablony povedeniia]. Vestnik MGU, ser.Filologiia. Moscow, 2002. №5, pp. 131–141.
7. *Lukashevich N. Iu.* 2004. Cognitive semantic analysis of predicates denoting human character traits. [Kognitivno-semanticheskii analiz predikatov, oboznachaiushchikh cherty kharaktera cheloveka]. PhD thesis. Moscow, 2004.
8. *Lukashevich N. Iu., and Kobozeva I. M.* 2011. Character nominations in ontological perspective. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoii Konferentsii "Dialog" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2011") (Bekasovo, 25–29 May 2011). Moscow, 2011. pp. 468–477.
9. *Martem'ianov Iu. S.* 1999. Metalanguages of sentence and text description.[Metaiazyki opisaniia predlozheniia I teksta]. Obrabotka tekst i kognitivnye tekhnologii. Mloscow, 1999. №3, pp. 124–129.
10. *Martem'ianov Iu. S., and Dorofeev G. V.* 1969. Logical inferences and eliciting relations between sentences in the text [Logicheskii vyvod i vyiavlenie sviazei

mezhdu predlozheniiami v tekste]. Mashinnyi perevod i prikladnaia lingvistika. Moscow, 1969, issue 12, pp. 36–60.
11. *Pen'kovskii A. B.* 1988 Adverb semantics and its representation in a dictionary [Semantika narrechiia i ee otrazhenie v slovare]. Slovarnye kategorii. Moscow, 1998, pp.53–59.
12. *Thomason R.,& Stalnaker R.* 1973 A semantic theory of adverbs. Linguistic Inquiry, 1973, v.4, №2, pp.195–220.
13. *British* National Corpus. Available at http://www.natcorp.ox.ac.uk/index.xml. (assessed at 10 February 2012).
14. *Longman* Dictionary of Contemporary English, 1995. Longman Dictionary of Contemporary English. Barcelona, 1995.
15. *Merriam* Webster Online. Available at http://www.merriam-webster.com (assessed at 10 February 2012)

CONSTRUCTING A LEXICO-TYPOLOGICAL DATABASE (FOR A STUDY OF PAIN PREDICATES)

Kostyrkin A. V. (languages@bk.ru),
Panina A. S. (panina-anna@yandex.ru),
Institute of Oriental studies, RAS

Reznikova T. I. (tanja.reznikova@gmail.com)
All-Russian Institute of Scientific and Technical Information, RAS

Bonch-Osmolovskaia A. A. (abonch@gmail.com)
Higher School of Economics

We present a database developed for lexico-typological study of expressions of pain (demo version available at http://orientling.ru/bolit/). Its design implements the non-relational, NoSql approach, where data is organized into a flexible tree not limited in size and depth. Linguistic annotation is placed directly into the text of example sentences and their translations, so that in effect the database is structured as an annotated corpus.
This formalism gives much freedom to both the developers in their task of annotating examples, and users in their queries, since it allows them to vary the level of detail according to how much information is available or needed.
Linguistic annotation includes tags for syntactic roles, some syntactic constructions and their components (relative clauses, light verbs, formal subjects, parts of compound words), morphological information (tags for case, number, aspect etc), as well as semantic tags specific to the domain of pain (semantic roles and types of metaphoric shift).

Key words: lexical typology, lexical semantics, database, NoSql, pain

1. Introduction

A prominent and promising method in lexical typology lately has been comparing semantically coherent word classes, rather than meanings of individual words, across a number of languages. This is the primary goal of the Moscow lexico-typological group, from the Aquamotion project [Maisak, Rakhilina 2007, Lander Y. et al. 2010] describing verbs of swimming and floating, to the present research into the vocabulary of pain [Britsyn, Rakhilina et al. 2009].

Invaluable in such comparative studies, especially when many languages are involved, are databases. They offer, first, a way to record any discovered rules and generalizations, to organize data and present it in a readily accessible form. Secondly, a database itself can be made into an analytical instrument, as it allows to search

for new correlations, verify hypotheses etc. The advantages of a database compared to a mere collection of raw data (even digitized) are obvious, and in direct proportion to its functionality.

Naturally databases have been employed in lexical studies for some time, yet none of the existing analogues suited the specific demands of our project.

WordNet [WordNet Lexical Database 2012], whose sheer scope makes it the flagship of lexical semantics representation, is ontologically rather than typologically oriented. With each language essentially independent from the others within its loose general framework, it is not suited for — nor, indeed, aimed at — the discovery of cross-linguistic phenomena. The same can be said about FrameNet [FrameNet Project 2012] and Lexicograph [Lexicograph Project 2012]. While each database provides an in-depth description for important areas of English and Russian lexicon respectively, neither was ever intended to encompass more than one language.

On the other hand, Anna A. Zalizniak's Catalogue of Semantic Shifts [Gruntov I. 2007, Zalizniak 2009] covers a number of languages and aims explicitly to serve typologically relevant observations, but the information it provides is limited, for the most part, to the source and target domains of the semantic shifts and does not detail any accompanying differences or similarities in argument realization and other syntactic behavior.

It is natural that a database's structure and content are determined by its purpose. However, even the single task of describing a relatively tight thematic class of predicates across several languages turns out to call for more than one database design. In fact, each thematic class appears to present a different set of requirements for its formal representation.

The Moscow lexico-typological group views a thematic class such as verbs of animal sound emission, motion in liquids, or pain in terms of an underlying system of oppositions that define relevant aspects of the situation, and, optionally, its relations to other domains. For aquamotion the main task was to find how this little domain itself is organized in each particular language, with such defining features as agency of the moving object (as in swimming vs. drifting) and directedness of motion (as in being carried by a current vs. floating in place). Verbs of animal sounds proved interesting first of all as a source of metaphoric shifts into other kinds of sound, speech and noise, and the database for the project's data was built around an elaborate classification of human, artificial and natural sound-emitters. Neither meta-structure could be used as is for pain predicates. For these we needed to accommodate secondary pain verbs recruited from non-algetic domains, and to fully capture this process we also had to pay more attention to the predicates' syntactic patterning. In other words, we had to design a database specifically suited to the project.

2. The domain of pain

What sets pain expressions apart from motion or sound is the way this domain, consistently across languages, is made up of a core of a few primary pain predicates and a greater number of secondary pain predicates.

Secondary predicates originate in one of the following domains: 1) combustion (e. g. English *my eyes* ***burn***) and related processes involving high temperature (e. g. German *glühen* 'glow with heat', Crimean Tatar *qajnamak* 'boil'); 2) deformation or destruction, in particular impact with sharp instruments such as blades or needles (e. g. Chinese *yāobù* ***cìtòng*** lit. 'the side pricks') and quasi-instruments such as claws or thorns (e. g. German *meine Augen* ***beißen*** lit. 'my eyes bite'); 3) motion, and deformation through motion (e. g. Ukrainian *nogi* ***krutit'*** lit. '[it] twists the legs'); 4) sound (e. g. Chinese *dùzi* ***jiào*** lit. 'the stomach screams') [Britsyn, Rakhilina et al. 2009].

One characteristic of a pain expression, therefore, is the type of metaphoric shift that created it, and we wanted to be able to keep examples of the original, non-metaphoric usage where applicable.

Syntactically primary pain verbs are statives, and as such tend to take a single argument interpreted as a theme or patient. Many of the secondary pain verbs, on the other hand, originate from transitive activities, and any changes in argument realization accompanying their shift into the domain of pain are important.

The basic situation of pain involves three participants: the experiencer, the affected body part, and a cause or stimulus, and being faced with a range of possibilities for syntactic realization of each of these arguments meant we needed to capture syntactic patterning in more or less detail, preferably including some morphological information. E.g., verbal aspect in languages such as Russian (imperf. *kolot'* vs. perf. *kol'nut'*) is involved in distinguishing continuous and momentary pain.

On the other hand, semantic classes are not really relevant to our task. For the body part and the experiencer they are trivial (the experiencer in our data is always human; while an animal experiencer is theoretically possible, the distinction of human vs. animal does not seem to bear any particular significance). For the predicates we decided not to subdivide the algetic domain into subdomains to represent specific kinds of pain as a medical phenomenon. Such classification would run a high risk of being arbitrary, since the experience of pain is highly subjective. Instead it is usually enough to know the stimulus and rely on our extralinguistic knowledge of the effect a given stimulus is likely to produce (e. g. the reaction of eyes to soap vs. bright light vs. strain). Accordingly we recognize the importance of collecting and classifying stimuli, while the most important feature of an algetic predicate is considered to be its domain of origin.

Our prospective database user, a semantic typologist, could need any of these types of data in any combination.

3. Annotation

One crucial decision that was made from the start was adopting the NoSql approach for database representation, opposite to the relational one. Within this approach linguistic data is structured not as a table with fixed number of fields but as a flexible tree not limited in size and depth.

Neither in retrieving nor in inputting data by this method does one have to struggle with a fixed set of obligatory features, some of which may be irrelevant to a given

sentence. In this formalism providing meta-information is a matter of applying tags directly to the words that warrant them, without pulling the original sentence apart. In essence, the database is organized as an annotated corpus of sample sentences.

This is a novel method for computational studies of lexical typology, and it offers incomparably greater freedom in both in the content (e. g., without any predetermined formal requirements on patterning, any given example sentence need not have all, or any particular subset, of the basic participants — experiencer, body part, stimulus — realized in any particular way), and in the annotation, which easily accommodates subsequent corrections and additions.

This ease of managing meta-information proved a particularly significant benefit. The project gradually came to encompass the data from more then twenty differently structured languages such as Korean, Japanese, Spanish, French, German, Czech, Ukrainian, Serbo-Croatian, Hindi, Chinese and Crimean Tatar; but even without additional languages the mere accumulation of data often led to substantial changes in annotation.

For example, initially we recorded patterns of argument realization in what was hoped to be an exhaustive list. We would tag the predicate with a case frame from the list, such as *s. d. o.* for taking a subject, a direct object, and a dative object, and repeat the *s.*, *d.* and *o.* as its arguments' tags. There were similar tags for verbs taking oblique objects, possessives and clauses.

It was not long before ellipsis transpired as a strong factor. Was the verb still to be tagged *s. d.o* when no dative argument, or no argument except the subject, was present in the sentence, just to show the predicate's combinatory potential? Including only examples with a full set of participants seemed impractical, as well as contrary to actual usage.

This and some other considerations eventually led us to stop tagging the predicate with its argument realization pattern at all, and mark argument realization where it was happening, i. e. directly on the arguments, in as much syntactic and morphological detail as necessary.

Incidentally this revision of the syntactic annotation helped to deal with the need to make generalizations over transitive verbs in ergative languages (formerly tagged *a. o.*) and nominative languages (tagged *s. o.*). The ergative subject tag *a.* was abolished, and arguments got separate tags for syntactic role and for case.

The predicate, on the other hand, was newly marked for being the pain predicate, and for part of speech to account for nominals (*pains and aches*) and adnominals (*an aching tooth* etc).

Relative constructions (as in French *J'ai la gorge qui brûle* lit. *'I have a throat that burns'*), anaphoric constructions, impersonal constructions with a formal subject, and compounds (such as *heartburn*) each also required a set of specific tags for their constituents.

Another major change happened with the list of metaphors for secondary pain verbs. The list grew with the addition of new languages, e. g. the metaphor of light was added for the Serbo-Croatian verb *sevati,* lit. '*shine*', which is used to describe rheumatic aches. But apart from merely expanding the list of metaphors, at one point we introduced some, if limited, internal structure, such as allowing the universal tag

for destruction *DESTR* to combine with the tag *SELF* for spontaneity (as in *My head is splitting*), or *INSTRUM* for instrumental action (as in *stabbing pain*).

The general principle behind the non-relational formalism is to include as much information as is available — for example, there are tags to mark a noun for oblique *and* a particular case such as ablative *and* gender *and* number, if we want, — but none of it is required; so that when some features are not applicable to the data (e. g. the sentence is in a language that has no grammatical gender or number) there is no need to tag them with default values, and when there is little information available the tags can also be minimal. Even the most essential tags for the participants of the algetic situation, their semantic roles — *exp* for experiencer, *bp* for body part, *s* for stimulus and *p* for the pain predicate itself, — are only present when the respective participants are present.

The same is true for the database user's side. Exactly the same mechanism serves to make a query as general or specific as needed, while freely combining semantic, syntactic and morphological information. Thus, a user can search for all sentences with a body part and an experiencer, or an experiencer realized as a subject, or as a subject in the nominative case only, or for a specific lexeme in any role, but only in the plural, etc — the flexibility is endless.

There is no need to resort to predefined constructs such as "transitive verb with a dative object". The user can easily describe it for themselves by requesting a combination of conditions — namely, that the pain predicate be a verb, with a direct object and a dative object present in the same sentence — and, furthermore, has an opportunity to relax (or tighten) these restrictions, which would not be possible if the syntactic pattern was a single feature.

4. Implementation

Examples as they are appear in the search results (available as a working demo version at http://orientling.ru/bolit/) consist each of a sentence in national writing and/or transliteration, followed by its Russian translation, literal and/or literary:

The data as the annotator sees it when inputting and editing examples include the same plus the markup. E.g., the German sentence in Fig.1, *Lisa tat vom lauten Reden der Hals weh 'Lisa's throat hurt her from speaking loudly'*, looks as follows:

de: [[*Lisa*/exp/D/DAT=ru:*Лиза*]] [[*tat*/LV.DELAT/PAST.SG <*tun*=ru:*делать*]]
[[*vom lauten Reden*/st/OBL/ABL=ru:*громкая речь*]]
[[*der Hals*/bp/S/NOM.SG=1]] [[*weh*/p/ADV.GENERIC=ru:*больно*]]
ru: (букв.) *Лизе горло делало больно от громкой речи*.
ru: *У Лизы от громкой речи заболело* [[горло=1]]
#' (Jentzsch, K., Ankunft der Pandora)

Here we can see, enclosed in double square brackets and supplied with tags, the main unit of description: participant groups. The order of tags is not relevant, but traditionally we place semantic roles first. Beside the experiencer, stimulus, body part

and pain predicate groups, there is one more group, not annotated with a semantic role — [[*tat*]], which is a purely syntactic part of the pain construction. The verb is marked *LV* for light verb, followed by its type *DELAT* 'do' (the other LV types are 'be' and 'have'). The pain predicate, adverbial *weh*, is marked for *GENERIC* type, meaning non-metaphoric, primary pain word.

Fig. 1. Search results for *'throat'*

Head words in some of the groups are also annotated with the word's dictionary form and Russian translation. This is done so that the sentence could be found, e. g., by searching for *tun* even though it only contains the inflected form *tat*. For *Hals* there is no need to place the Russian equivalent directly into the group in the source sentence, because, as Fig.1 proves, the example is searchable via the translation sentence. The Russian equivalent is only of secondary importance, because the source lexical item is identifiable by its dictionary form.

Here also can be observed an experimental feature for better readability: *der Hals* in the source sentence and its equivalent *горло* in the translation are followed by an identifying number, which causes them to display underlined in a similar color and thus ties them together structurally and visually. The same notation can be used when one source word corresponds to several, not necessarily adjacent, words in the translation or vice versa. For example, to underline the correspondence of *tat* ... *weh* lit. '*did* (her) *pain*' to *заболело* 'hurt' we just need to add another number to their respective groups:

de: [[*Lisa*/exp/D/DAT=ru:*Лиза*]] [[*tat*/LV.DELAT/PAST.SG <*tun*=ru:*делать***=2**]] [[*vom lauten Reden*/st/OBL/ABL=ru:*громкая речь*]] [[*der Hals*/bp/S/NOM.SG=1]] [[*weh*/p/ADV.GENERIC=ru:*больно***=2**]]
ru: *У Лизы от громкой речи* [[*заболело***=2**]] [[горло=1]]

There is a mechanism to check the notation for formal correctness when a sentence is added or edited. Besides errors which preclude an internal representation of the sentence from being formed (mostly syntactic, such as missing brackets), some situations which do not prevent the sentence from being parsed and saved into the database also cause a warning message to appear. One such message can be seen in Fig.2, warning about an unknown tag. In this case the tag is misspelled, but if it really was new and for some reason not yet on the list, it would not have prevented the example from being added.

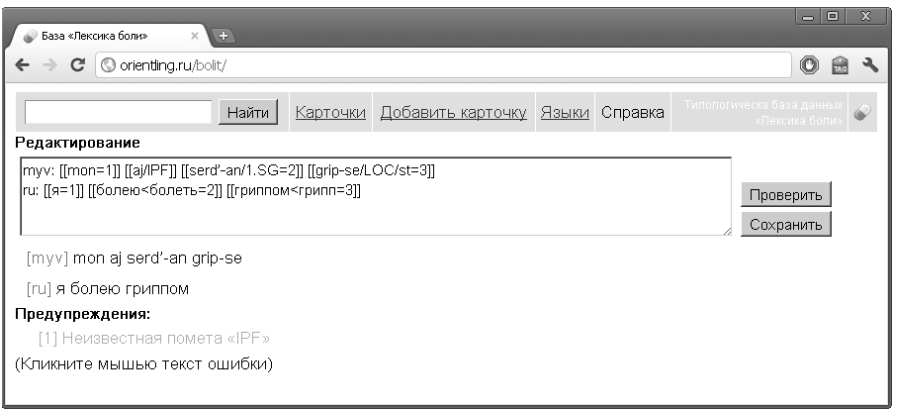


Fig. 2. Edit window displaying a warning message for an unknown tag

To prevent tags from proliferating, annotators cannot add tags. Suggestions for any changes in the annotation are submitted to the database administrator; they are then discussed by all developers, and adopted as a joint decision.

5. Conclusion

In conclusion it can be said that the database of pain expressions, aside from its role as such, is also an attempt to build a flexible, versatile instrument for lexicotypological studies in general.

The purely grammatical notation, syntactic and morphological, is independent of the algetic domain and can be used elsewhere as is, or modified and expanded to accommodate more grammatical categories and syntactic constructions as new languages or tasks demands. The three semantic roles currently in use are, naturally, specific to the domain of pain and will have to be replaced with what is appropriate to the new frame describing the new semantic class. Similarly the list of metaphors, if a need arises to study words from the perspective of their origin in other domains, will need to be adjusted.

What we consider most significant, however, is the approach itself, embodied in our choice of non-relational formalism.

Not prescribing how a construction should be structured, not imposing predefined theoretical constructs onto data, but describing what is actually present in the text while allowing generalizations to naturally emerge as patterns of co-occurrence, it lets both the database compilers and users choose freely the amount of information to provide or demand, vague or specific, depending on their needs at the moment.

References

1. *Britsyn V. M., Rakhilina E. V., Reznikova T. I., Iavorskaia G. M.* eds. (2009) Concept of Pain in the Light of Typology [Kontsept bol' v tipologicheskom osvesche-nii]. Kiev.
2. *Gruntov I. A.* «The Catalogue of Semantic Shifts»: a Database for the Typology of Semantic Evolution. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2007" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2006"]. Bekasovo, 2007.
3. *FrameNet* Project (2012), available at https://framenet.icsi.berkeley.edu/fndrupal/
4. *Lander Y., Maisak T., Rakhilina E.* (2010) Domains of aqua-motion: a case study in lexical typology, in Motion Encoding in Language and Space, Oxford University Press, Oxford.
5. *Lexicograph* Project (2012), available at http://lexicograph.ruslang.ru/ (in Russian)
6. *Maisak T. A., Rakhilina E. V.* eds. (2007) Verbs of AQUA motion: lexical typology [Glagoly dviženija v vode: leksičeskaja tipologija], Indrik, Moscow.
7. *Zalizniak Anna A.* On the notion of semantic shift [O ponjatii semanticeskogo perehoda]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2009"]. Bekasovo, 2009, pp. 107–112.
8. *WordNet* Lexical Database (2012), available at http://wordnet.princeton.edu/

THE RUSSIAN EMOTIONAL CORPUS: COMMUNICATION IN NATURAL EMOTIONAL SITUATIONS

Kotov A. (kotov@harpia.ru)
National Research Centre "Kurchatov Institute", Moscow, Russia

Budyanskaya E. (davemonty2@gmail.com)
Research Computing Center, M. V. Lomonosov Moscow State University, Moscow, Russia

The Russian Emotional Corpus (REC) includes annotated video recordings of natural communication in tense emotional situations: oral university exams and talks between clients and officers at a municipal office regarding utility bills. Annotation of the corpus describes speech (text, syntagmatic structure, speech acts, face threatening acts and the cases of irony), facial expression, gaze direction and hand gestures. In the recorded emotional situations informants show diverse emotional cues and numerous rational and emotional communicative strategies. The annotation allows us: (a) to search for a specific cue (like smiles or squints) and describe its usage and functions in natural communication, or (b) to search for specific patterns — like behavior markers of hesitation or facial cues at the end of utterances with face threatening acts. These observations help us to animate emotional computer agents, which receive semantic trees at their input, simulate rich emotional dynamics and produce semantic trees, ready-made phrases and gestures for the output.

Key words: multimodal corpus, emotional expression, emotional communication, emotional agents

1. Introduction

Linguistic corpora recently became an important base for theoretical studies as well as for training and verification of applied linguistic tools. Widespread text corpora are combined with audio corpora and multimodal corpora, allowing the study of speech, phonetics, facial expression, gestures and other aspects of real face to face communication. We collect and annotate the Russian Emotional Corpus (REC) with video records of natural emotional communication. The corpus is mainly aimed at the creation of emotional computer agents, which recognize emotional cues in communication and simulate emotional behavior in a dialogue.

1.1. Usage and types of emotional corpora

The development of entertainment technologies, computer agents (like game characters, interface assistants) and mobile robots as well as attention to emotional

aspects in all types of communication stimulate the development of emotional corpora. Emotional multimodal corpora are represented by (a) the observed public data: records from mass media [Abrilian, Devillers et al., 2005], (b) performed data: where actors have to perform given emotions [Bänziger, Scherer, 2007], and (c) experimental data: where the emotional arousal is induced in experimental situations [Zara, Maffiolo et al., 2007].

At the same time, this data is still different from real face to face interaction. The behavior of informants is restricted (induced) by the situation (environment) of publicity, acting or experiment. As a result this data may miss a number of natural emotional phenomena: conflicts between internal motivation, goals and suppressed emotions — on one side, and strategies of politeness, necessity to simulate "polite" or "desirable" emotions (like etiquette smiles) — on the other side.

This requires the collection of real emotional data, as for example, made in [Scherer, Ceschi, 1997]. This is a difficult task for several reasons: (a) during communication subjects should remain in front of a fixed camera (unlike in spoken corpora), (b) camera interferes with communication, (c) we cannot manipulate the situation, for example asking a salesman to frustrate the buyer (unlike in experimental cases), (d) severe legal limitations apply to all public cases of video records, and (e) ethical and legal considerations force us to limit access to the corpus.

At the same time, we may get ecologically valid data (subject to possible influence of (b)) containing diverse communicative strategies and complicated emotional patterns, suitable for studies and modeling.

1.2. Emotional computer agents

Natural emotional data is important for dialogue support systems, which should "understand" clients in natural situations, and for the development of software agents with rich emotional architectures [Rehm, André, 2008].

For example, Max computer agent uses different incoming phrases to change its emotional state and to balance between "good" and "bad" mood, producing different expressive cues and phrases for each local state [Becker, Kopp, Wachsmuth, 2004]. Greta computer agent uses two different (conflicting) emotional arousals (internal emotions and masking superficial emotion) to construct compound patterns of facial expression — "blended emotion" [Ochs, Niewiadomski, et al., 2005], as observed in mass media corpus [Abrilian, Devillers, et al., 2005; Maglogiannis, Karpouzis, et al. 2006].

In our studies we develop emotional computer agents, which receive semantic trees at their input, simulate emotional dynamics and produce semantic trees, ready-made phrases and gestures in BML format for the output. The agents are designed to simulate complicated emotional dynamics in communication, in particular:

(i) emotional oscillation: an incoming event or phrase may simultaneously activate numerous responses (like anger, confusion, desire for reconciliation, etiquette regulations, rational intention to solve the situation etc.), which compete for the output, forcing the agent to oscillate between several short communicative positions or emotions (*microstates*) in its behavior [Kotov, 2007];

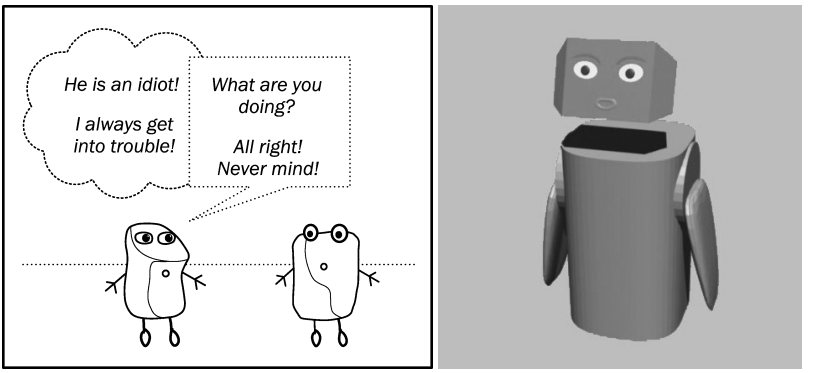


Fig. 1. 2D and 3D emotional computer agents

(ii) top-down emotional effects (as discussed in [Clore, Ortony, 2000]): in a negative state the agent treats incoming stimuli as negative, changing the semantic representation of an incoming phrase or event — if we 'touch' the agent, it believes, we 'beat him' (if in a negative state), or that we cherish him (if in a positive state) [Kotov, 2007];

(iii) hypocrisy and sarcasm: if we 'beat' the agent, forcing its negative arousal, he may suppress the immediate expression and automatically select the best positive reaction (like *It's good you have paid attention to me!*) to use it in a sarcastic way [Kotov, 2009].

(iv) switches in communication: the agent may address different parties around — the main counteragent (the person, who has just committed an action or was affected by the agent's action) and observers with different levels of the agent's confidence (friends or strangers); the agent may also convey its "thoughts" (too aggressive to too sincere responses) to a special callout.

As the architecture is based on a number of conflicting tendencies, it should be well balanced in order to produce believable compound patterns. To observe and study naturally appearing complicated patterns we require a video corpus of real behavior in tense emotional situations.

2. Principles of collection

We collect records of emotional situations, where the investigated party (student, client) has very strong motivation to communicate and achieve his goal. Obstacles, that appear, may force internal conflicts, numerous expressive cues and usage of different influence strategies.

During the communication one of the parties (examiner, client) gives the instruction, and the other party (student, officer) should perform in order to satisfy the opponent. From this point of view we investigate two symmetrical situations, similar to future interactions with mobile robots and dialogue systems, which should fulfill human requests or otherwise satisfy the opponent through the emotional interaction.

The legislation obliged us to inform the recorded party and respect his privacy. We have turned off the camera in the cases of refusal (3 cases during the collection of the student corpus). We also gave an obligation not to publish or disclose the records, limiting their usage only to scientific purposes.

2.1. Oral university exams

Oral university exams have been widely replaced by written tests to eliminate the contribution of personal influence and impressions to the final score. This "undesirable" emotional interaction was exactly the object of our interest in the collected records. This corpus section contains 295 records of university exams and tests (length from 26 sec to 60 min, average 6 min, total length 29,5 hours; 5 humanitarian university courses, 2 universities). Each record includes full communication with one student: video of the student and audio record of the conversation. 194 records are annotated to the present moment.

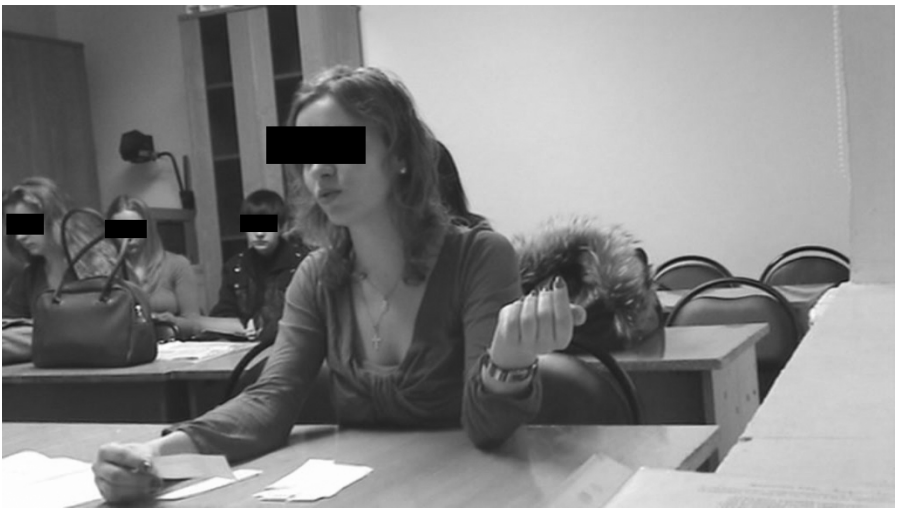


Fig. 2. Exam situation sample frame

2.2. "One window" municipal service

"One window" municipal service interacts with clients on all the questions regarding provided community facilities (water, central heating), issued bills, residential rental, discounts etc. REC contains test series, recorded in June 2009 (4 days) and the main series, recorded in the middle of October 2009 (15 days, 2,5 weeks at 6 working days). Corpus covers the full working time of the window (average 8 hours per day for the main series). The total recorded time (146 hours, including idle time) was

divided into 510 records, where each record includes full communication with one subject (length from 5 sec to 30 min, average 3,8 min, total length 32 hours). The records include video of a client and full audio of the conversation. For this corpus section only microstates are annotated to the present moment.

Fig. 3. "One window" sample frame

Many factors show, that the influence of camera was greatly reduced: in the exam corpus some students tried to cheat when an examiner left the room, in "one window" corpus only a few subjects noted the camera (not only the announcement about video recording).

3. Principles of annotation

The main goal for the annotation was to annotate expressive cues, specific for emotional interaction and distinguishing current emotional examples from imaginary steady and non-emotional behavior. We had to provide (a) reliable and simple superficial annotation, suitable for different search and pattern extraction tasks, and (b) specific functional annotation, oriented to our research purposes.

The corpus is annotated with the help of ELAN annotation software[1]. The annotation is divided into "tiers" and describes speech, facial expression and hand gestures, as shown in Table 1. We also annotate rapid changes in expressive or communicative strategies (changes in microstates) to rely on these examples during modeling of the agent's behavior. Head movements and body postures are not yet annotated.

[1] ELAN annotation software is developed at Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. Software URL: http://www.lat-mpi.eu/tools/elan/

Annotation scheme is always a compromise between accuracy (e. g. annotation of action units or allophones) and simplicity (e. g. annotation of cues and words/phrases). We follow the second strategy: we try to provide a simple index to a wide number of cases and annotate cues, which may be significant for emotional interaction, taking into account, that minor actions, constituents of each cue, can be studied later. To reduce the redundancy we join tiers for (a) eyes, eyebrows, nose and gaze and (b) lips, jaw and tongue — we consider, that mainly one cue (from our list) may be expressed in each of these areas and other units usually follow the expression (e. g. when subjects widely open their eyes, they usually move eyebrows up); although in this annotation we lose some cues, where eyes and eyebrows act separately (e. g. closing eyes and moving eyebrows up).

The records include diverse, indefinite and incomplete movements and gestures, so annotation of hand gestures has four independent parameters: active_organ, passive_organ, mode and trajectory (as in Table 1). This allows the description of non-standard, incomplete and partial gestures.

Annotation is made by one annotator, checked by a supervisor and stored on a version control server.

Table 1. Annotation scheme of the Russian Emotional Corpus

Group	Tier	Comment	Annotation type
Text	text1	Text of the main participant	Free text annotation, divided by speech segments
- " -	text_stru	Syntagmatic structure	Describes syntagmatic position of the current speech segment
- " -	text_comgoal	Illocutionary force	Indicates communication goal or illocutionary force
- " -	text_fta	Face threatening acts	Indicates possible face threatening acts and politeness strategies
- " -	text_irony	Irony	Indicates the cases of irony or explicit role-playing
- " -	text2	Text of the main addressee	(the above annotation tiers are duplicated for this line)
- " -	text3	Text of all other participants	- " -
Eyes	eyes	Eye, eyebrows and nose movements, specific gazes	Fixed list: *looks aside, looks up* (gazes are considered as expressive cues, continuous gaze direction is not annotated)*, opens* [eyes] *wide, squints, closes, winks, rises eyebrows, frowns, wrinkles nose, blinks rapidly, other*

Group	Tier	Comment	Annotation type
Mouth	mouth	Mouth movements	Fixed list: *licks lips, smiles, laughs, purses lips, chews* (moves lower jaw), *compresses lips, sucks in lips, bites lip, grins, moves lips* (as if silently speaking), *coughs, clicks, spits, inhales, exhales, opens mouth, other*
Hands	active_organ	Active organ for gestures	Fixed list: *single finger, fingers, palm, fist* (or back of the hand), *arm, adapter, other*
- " -	passive_organ	Passive organ or object for gestures	Fixed list: *none, hair, elbow, forehead, eyes-eyebrows, nose, cheek, mouth-lips, ear, chin, fingers, palm-hand, fist* (or back of the hand), *arm, own clothes or body, an object, human, environment* (immobile object, like table or wall), *other*
- " -	mode	Mode of operation	Fixed list: *manipulates, points, counts, iconic* (iconic gestures, manipulation with imagined objects), *touches, knocks, rubs, scratches, waves, demonstrates* (keeping an object with hands), *closes* (e. g. mouth with a hand), *supports, crosses* (fingers, palms, arms), *other*
- " -	trajectory	Trajectory of operation	Fixed list: *fixed* (cue has no trajectory), *to an object, to the addressee, away* (e. g. waving with hand away from the body), *towards the body, up* (e. g. scratches the nose upwards with a finger), *down, in discourse space* (manipulations in an imaginable space), *in the environment* (e. g. arranges objects on the table), *round, excursion* (preparation), *recursion* (release), *reverse* (movement to and from, e. g. when hesitating), *standby, fragment* (incomplete gesture), *other* (e. g. fiddles with clothes)
Micro-states	microstates	Microstates and sequences of microstates	Keywords for microstates — units and markers for the developed computer agent architecture
- " -	m_phases	Phases of microstates	Each expressive phase is marked as a separate annotation

4. Patterns of emotional interaction

Emotional communication in REC has certain distinctions: (a) informants show numerous and compound reactions on incoming stimuli, (b) informants try different strategies (both rational arguments and emotional influence) in order to get to their goal in communication, and (c) in both cases informants may show substitutive expressive means — substituted gestures and simulated emotions.

4.1. Multiple reactions

Subjects may balance between different speech strategies, mixing cues of their own arousal with required performance strategies. In the following example (Table 2) a student (female) gives a definition for "zeugma" and hesitates during the answer phrase, showing different speech and facial hesitation cues (record 20081226-zhurn-a1, 04:20).

The following dialogue in "one window" service shows switches in communication, when the client announces his "thoughts" in the second phrase, and changes in communication strategies, when the client after an emotional appraisal moves on to discuss a rational solution:

< The client (female) wanted to get a statement of her residential account, but forgot her passport >
Officer: *Only with passport.*
Client: *So only with passport, yeah?* < looks above the window > *Well, I'm not a lucky one.* < looks in the window > *But you're open every day, right?*
(clip 20091022-a03)

These examples constitute a valuable material on how to construct and balance different reactions (calling expressive cues or speech strategies) for an emotional computer agent.

4.2. Multiple strategies

In addition to numerous strategies of politeness, described in [Brown, Levinson, 1987], people show diverse emotional strategies to influence the addressee. In the exam section students (female) may start to smile and flirt with the examiner (male) after a wrong answer in order to smooth over the bad impression and achieve benevolence through the emotional interaction. Students and clients may simulate and exaggerate cues for pain, tiredness, frustration, nervousness in order to show, that the situation is negative and should be solved by the opponent: examiner or municipal officer [Kotov, 2011]. This may be combined with other strategies: attempts to command the opponent and to make him follow the applicable rules. In particular, in 20081225-zhurn-b3 a student (female) during a 18 sec fragment

consecutively (a) tries to provoke sympathy and come to take the test another time, (b) tries to provoke indulgence by showing, she does not feel/understand very well, (c) rationally asks to specify the task, (d) accuses the opponent, indicates, that the task is not precise, (e) laughs in order to reduce a negative impression from her past words. In the intense exam situation here the influence strategies are changed without correct coordination every 3,6 sec (on average), which makes this example a valuable base to design computer agents, able to enumerate possible emotional strategies in communication.

Table 2. Speech protocol with facial expression

Speech	**Mouth**	**Eyes**
Zeugma is…		
	licks lips	looks aside < right > rapidly blinks
when…		looks aside < left >
	opens mouth	
	compresses lips; clicks	looks aside < down >
ep… ekh…		looks aside < right >
some word…		<looks at the examiner> rapidly blinks
mostly the verb		looks aside < right > looks aside < down >
can be omitted.		<looks at the examiner>

5. Discussion

The corpus gives us numerous examples of emotional expressive cues and strategies, which are difficult to observe in other types of data.

Search function, provided by ELAN annotation software, allows to look for a specific cue or a combination of cues (possible expressive pattern): e. g. 'raises eyebrows and compresses lips' (n = 47) or 'compresses lips and immediately after licks the lips' (n = 94). By browsing the search results we can (a) classify the superficial expressive types for a cue, e. g. types of smiles (mouth open, mouth closed etc.) or clicks (bilabial, labiodental, coronal etc.), and (b) classify contexts and functions for a specific cue or expressive pattern, e. g. position of smiles in speech, functions of smiles. We also can check, if the combination of cues constitutes a steady expressive pattern with some specific function, or if the two cues just co-occur on the timeline. This allows the design of an expressive library for a computer agent with "grammar" for each expressive cue, which further may allow the establishment of functional annotation for the superficial cues in corpus.

Observation of compound examples suggests the architecture of the computer agent, which should be sufficient to construct dynamically these examples during its

operation. In particular, examples with numerous expressive cues and numerous influence strategies require the architecture, where different reactions are activated for a given stimulus (or goal) and compete for the output (speech and animation of the agents figure).

6. Conclusions

The corpus provides ecologically valid data with annotation and can be used for a variety of scientific and applied purposes:

- analysis of functions and "grammar" of facial expression and gestures;
- analysis of communication strategies: humor, politeness strategies, exploitation of emotions ("pull" emotions);
- extension for the theories of conversational analysis and design of automatic dialogue systems;
- automatic construction of "speaker's intent" or "the theory of mind", when a computer system interprets observed expressive cues in order to detect the internal state and intent of the speaker;
- automatic pattern extraction;
- training of automatic classifiers (e. g. computer neural networks) in the area of dialogue simulation;
- database of complicated expressive patterns for cartoon animators and actors;
- training base for state and corporate officers, interacting with clients; optimization of corporate client services.

In our case, the observations made on the basis of the corpus, allow the extraction of expressive patterns for emotional computer agents, and the design of the internal architecture of the agents to simulate the observed compound reactions, involving internal emotional arousal as well as surface etiquette and simulated emotional expressions.

References

1. *Abrilian S., Devillers L., Buisine S., Martin J.-C.* EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. HCI International 2005. Las Vegas, USA, 2005.
2. *Bänziger T., Scherer K. R.* (2007) Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The GEMEP Corpus, in Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science 4738. Springer-Verlag, Berlin, Heidelberg, pp. 476–487.
3. *Becker C., Kopp S., Wachsmuth I.* (2004) Simulating the Emotion Dynamics of a Multimodal Conversational Agent, in Affective Dialogue Systems, Lecture Notes in Computer Science 3068, Springer-Verlag, Berlin, Heidelberg, pp. 154–165.

4. *Brown P., Levinson S. C.* (1987) Politeness : Some Universals in Language Usage, Cambridge.
5. *Clore G. L., Ortony A.* (2000) Cognition in Emotion: Always, Sometimes, or Never? in Cognitive Neuroscience of Emotion, Oxford Univ. Press, pp. 24–61.
6. *Kotov A.* (2007) Simulating Dynamic Speech Behaviour for Virtual Agents in Emotional Situations, in Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science 4738. Springer-Verlag, Berlin, Heidelberg, pp. 714–715.
7. *Kotov A.* Accounting for irony and emotional oscillation in computer architectures. International Conference on Affective Computing and Intelligent Interaction ACII 2009. Amsterdam, IEEE, 2009, pp. 506–511.
8. *Kotov A.* Types of Simulated Emotional Expressive States in the Russian Emotional Corpus Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Po Materialam Ezhegodnoi Mezhdunarodnoi Konferentsii "Dialog" [Computational Linguistics and Intellectual Technologies: Materials of the International Conference "Dialog"]. Bekasovo, 2011, pp. 315–324.
9. *Maglogiannis I., Karpouzis K., Bramer M., Martin J.-C., Caridakis G., Devillers L., Karpouzis K., Abrilian S.* (2006) Manual Annotation and Automatic Image Processing of Multimodal Emotional Behaviors in TV Interviews. Artificial Intelligence Applications and Innovations, Vol. 204, pp. 369–377.
10. *Ochs M., Niewiadomski R., Pelachaud C., Sadek D.* (2005) Intelligent Expressions of Emotions, in Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science 3784, Springer-Verlag, Berlin, Heidelberg, pp. 707–714.
11. *Rehm M., André E.* (2008) From Annotated Multimodal Corpora to Simulated Human-Like Behaviors, in Modeling Communication with Robots and Virtual Humans, Lecture Notes in Computer Science 4930, Springer-Verlag, Berlin, Heidelberg, pp. 1–17.
12. *Scherer K. R., Ceschi G.* (1997) Lost luggage emotion: A field study of emotion-antecedent appraisal, Motivation and Emotion, Vol. 21, pp. 211–235.
13. *Zara A., Maffiolo V., Martin J.-C., Devillers L.* (2007) Collection and Annotation of a Corpus of Human-Human Multimodal Interactions: Emotion and Others Anthropomorphic Characteristics, in Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science 4738, Springer-Verlag, Berlin, Heidelberg, pp. 464–475.

РЕГИОНАЛЬНЫЕ ПРЕДПОЧТЕНИЯ В ПРОИЗНОШЕНИИ НЕКОТОРЫХ РУССКИХ И ЗАИМСТВОВАННЫХ СЛОВ

Краснова Е. В. (krasnova@speechpro.com),
Смирнова Н. С. (nsmirnova@speechpro.com)

ООО «Центр речевых технологий», Санкт-Петербург, Россия

Данное исследование посвящено изучению вариативности в произношении тринадцати слов, допускающих несколько вариантов произношения в русском языке. Исследовалась речь 240 информантов из 12 городов России. Интересным представлялось установить, зависит ли выбор того или иного варианта произнесения слова от региональной принадлежности говорящего, имеет ли место внутридикторская вариативность в выборе вариантов произнесения, влияют ли на произносительные предпочтения иные факторы (в частности, возраст).

В результате проведенного исследования однозначных региональных предпочтений выявлено не было: следует скорее говорить о более или менее выраженной тенденции к предпочтению того или иного варианта. Заслуживает внимания тот факт, что в ряде случаев наиболее частотными оказались варианты произношения, не рекомендованные орфоэпическими словарями. Также следует отметить, что в речи нескольких информантов встретились разные варианты произношения одного и того же слова.

Полученные данные о частотности произносительных вариантов, статистике их употребления в разных регионах, а также о внутридикторской вариативности в произношении могут служить для целей криминалистики, а также быть учтены при составлении орфоэпических словарей.

Ключевые слова: региональная диагностика, диагностический признак, вариативность, вариант, произношение, норма

1. Введение

Изучению диалектной вариативности русской речи в области фонетики, морфологии и синтаксиса посвящено множество исследований. Среди наиболее авторитетных можно назвать, например, работу К. Ф. Захаровой, В. Г. Орловой «Диалектное членение русского языка» [5], а также другие работы.

Как свидетельствуют результаты относительно недавних полевых исследований [6], до сих пор в речи жителей некоторых регионов России (преимущественно пожилых) сохраняются яркие диалектные черты. Одновременно с этим ведутся и специальные исследования, посвященные городской речи (т. н. городским социолектам), которая, хотя и утратила во многом исконный диалектный

колорит, тем не менее, все еще сохраняет в той или иной мере языковую специфику, свойственную местным говорам [4]. Часто отмечается, что речь современных городских жителей изобилует т. н. просторечными особенностями, общими для различных регионов и в определенных аспектах сближающимися с литературной разговорной речью. Современный этап развития русского языка характеризуется усиленным проникновением просторечия в языковую норму.

Данное исследование посвящено изучению региональной вариативности в произношении некоторых слов, допускающих несколько вариантов произношения в русском языке. Исследовалась только речь городских жителей, т. е. людей, в речи которых отсутствуют яркие диалектные черты в области морфологии и лексики, хотя в большей или меньшей степени может сохраняться диалектная фонетическая специфика. Анализируемые произносительные особенности, в большинстве своем, не связаны с исконными диалектными особенностями, тем более что часть из исследуемых слов — иноязычные заимствования. Как правило, выбор того или иного варианта произнесения в той или иной мере допустим современной орфоэпической нормой — в словарях один из вариантов приводится как основной, другой помечается как дополнительный или устаревший, хотя в разных словарях. В связи с этим интересным представлялось установить, зависит ли выбор того или иного варианта произнесения слова от региональной принадлежности говорящего, имеет ли место внутридикторская вариативность в выборе вариантов произнесения, могут ли на произносительные предпочтения влиять иные факторы (в частности, возраст).

Результаты исследования дают представление о произносительной тенденции в отношении ряда слов на современном этапе, могут быть учтены при составлении орфоэпических словарей, а также использованы при проведении экспертного сравнения речевых образцов.

2. Материал и методика

Исследование проводилось на материале звукозаписей речи жителей двенадцати регионов России. По характеру произносительных особенностей, которые прослеживаются в речи информантов, регионы можно классифицировать следующим образом:

1. Регионы, в речи жителей которых сохраняются черты исконно русских говоров:

- *Архангельск* (черты севернорусских говоров);
- *Великий Новгород* (черты севернорусских говоров);
- *Нижний Новгород* (черты севернорусских, среднерусских говоров);
- *Саратов* (черты севернорусских, южнорусских говоров);
- *Псков* (черты западных среднерусских говоров);
- *Воронеж* (черты южнорусских говоров);

2. Регионы, в речи жителей которых по причине русско-национального двуязычия проявляется фонетическая интерференция:

- *Казань* (влияние татарского языка);
- *Ижевск* (влияние удмуртского языка);

3. Регионы Сибири, на речь жителей которых оказывают влияние старожильческие говоры:
- *Иркутск* (также возможно влияние бурятского языка);
- *Барнаул* (также возможно влияние алтайских языков);

4. *Оренбург* (т. н. территория позднего заселения, преимущественно черты южнорусских говоров);

5. *Воркута* (наиболее поздно заселенная территория (среди исследованных), преимущественно черты севернорусских говоров, также возможно влияние языка коми).

В рамках исследования слуховым методом анализировались звуковые записи спонтанной речи и чтения двадцати дикторов из каждого региона. Все информанты — мужчины в возрасте от 20 до 60 лет. Требования к информантам в данном случае не были столь же строгими, как при выявлении диалектных особенностей. В подавляющем большинстве своем это коренные жители исследуемых регионов, родители которых также являются местными жителями. Однако поскольку ряд регионов имеет сложную историю заселения и высокую миграционную динамику (например, Барнаул), среди информантов были и те, кто прожил в целевом регионе только последние 20 лет своей жизни.

Исследовалось произношение следующих 13 слов: *дождь*, *дожди*, *уезжаем*, *отсюда*, *компьютер*, *сервис*, *прогресс*, *судоверфь*, *твердый*, *две*, *дверь*, *свет*, *свежий*. Часть исследуемых слов присутствовала в тексте, который предлагалось прочесть информантам, а часть прозвучала в ответах на заданные информантам вопросы.

Каждое из исследуемых слов может быть произнесено как в «мягком», так и в «твёрдом» варианте. При этом в разных орфоэпических словарях, как уже указывалось выше, статус того или иного варианта произнесения может различаться. Ниже приводим варианты произношения из словаря под редакцией Р. И. Аванесова [9]:

Дождь	до[ʃ':] *и* до[ʃt']
Дожди	до[ʒ':]и *и* до[ʒd]и
Уезжаем	уе[ʒ':]аем *и доп.* уе[ʒ:]аем
Отсюда	*доп.* от[su]да
Компьютер	компью[te]р
Сервис	[se]рвис
Прогресс	*доп.* прог[re]сс
(Судо)верфь	вариантов нет
Твердый	[t'v']ердый *и* [tv']ердый
Две	[d'v']е *и* [dv']е
Дверь	[d'v']ерь *и* [dv']ерь
Свет	[s'v']ет *и* [sv']ет
Свежий	[s'v']ежий *и* [sv']ежий

Как видно, для большинства слов варианты произнесения со смягчением и без него являются равноправными (авторы поясняют, что порядок следования рекомендуемых вариантов в данном случае незначим). Для слова (судо)верфь авторы не приводят транскрипционных обозначений, но исходя из приоритетных произнесений других слов, можно предположить, что основным вариантом будет [r'f'] с регрессивной ассимиляцией по мягкости. Для слов «уезжаем», «отсюда», «прогресс» вариант произнесения с твердым согласным является допустимым, в то время как основной вариант предполагает смягчение согласного. Для заимствований «компьютер» и «сервис», напротив, вариант со смягчением не приводится даже в качестве допустимого.

Также для исследования использовался «Словарь трудностей русского произношения» М. Л. Каленчук и Р. Ф. Касаткиной [7]. Оценка нормативности вариантов произношения исследуемых также будет приводиться в статье.

Запись материала, содержащего произнесения исследуемых слов, проводилась на диктофон ГНОМ-2 (ЦРТ) с частотой дискретизации 48 кГц, частотой квантования 16 бит.

Далее был проведен слуховой анализ фонограмм с целью определения, какой вариант произношения того или иного слова предпочтителен для того или иного информанта.

Слуховой анализ проводился с использованием звукового редактора SIS II (ЦРТ) — применялось «фокусное» посегментное прослушивание материала, а также, при необходимости, привлекались данные акустического анализа (значения формантных частот).

Результаты слухового анализа фиксировались в сводной таблице и затем подсчитывались в абсолютных и относительных единицах.

3. Результаты

3.1. Варианты произношения слов «дождь», «дожди»

В русском языке существуют два варианта нормы: петербургский и московский. Считается, что произношение слов «дождь и «дожди» с бифонемным сочетанием, то есть *do[ʃt']*, *do[ʒd']u*, характерно для петербургского варианта произносительной нормы, а произношение *do[ʃ':]* и *do[ʒ':]u* — для московского варианта. Следовательно, оба варианта произношения являются нормативными. Авторы словарей [7, 9] оценивают об варианта произношения как равноправные. Л. А. Вербицкая [3] отмечает любопытную деталь: произнесение [ʃt'] и [ʒd'] в словах «дождь» и «дождя» вместо [ʃ':] и [ʒ':] у москвичей встречается теперь чаще, чем у петербуржцев: соответственно в 74 % и 86 % случаев против 69 % и 81 % у петербуржцев, хотя произнесение твердых [ʃ] и [ʒ] в этих словах было одной из наиболее характерных черт старого петербургского произношения. Отметим также, что как в Петербурге, так и в Москве сегодня ведущим является бифонемное произношение, что пока не нашло отражения в наиболее известных орфоэпических

словарях. При этом переход к бифонемному варианту, по-видимому, легче происходит в форме множественного числа, о чем также свидетельствуют эти данные.

Аналогичная тенденция была отмечена и нами в процессе исследования: далеко не все информанты, речь которых тяготеет к московскому варианту произношения, то есть те дикторы, которые произносят *до[ʃ:']* («дождь»), слово «дожди» произнесли с долгим мягким [ʒ:']. Так, если слово «дождь» с бифонемным сочетанием [ʃt'] произнесли лишь 75% дикторов (что хорошо согласуется с результатами для московской нормы, полученными Л. А. Вербицкой), то слово «дожди» почти 100% информантов произнесли в соответствии с петербургским вариантом нормы. В речи 7,5% информантов реализовался своего рода промежуточный вариант произношения слова «дождь» — с бифонемным сочетанием с мягким [ʃ't'] (*до[ʃ't']*).

В данном случае невозможно говорить о зависимости предпочтений в произношении от региональной принадлежности информанта, но можно отметить различную степень стабильности регионов в выборе того или иного варианта. Так, наибольшая стабильность в произношении слова «дождь» с бифонемным сочетанием наблюдается в Оренбурге, Казани и Нижнем Новгороде: в этих регионах ни один информант не произнес *до[ʃ':]*.

Таблица 1. Статистика вариантов произношения слова «дождь»[1]

Вариант произношения	Количество реализаций в %
до[ʃ t']	75
до[ʃ':]	15
до[ʃ't']	8
более одного варианта	2

Таблица 2. Статистика вариантов произношения слова «дожди»

Вариант произношения	Количество реализаций в %
до[ʒd']и	99
до[ʒ':]и	1

Таблица 3. Встречаемость вариантов произношения слова «дождь» в регионах в %

	Реализации с бифонемным сочетанием	до[ʃ':]	Реализованы оба варианта
Воркута	70	30	0
Оренбург	**100**	**0**	**0**
Барнаул	90	10	0

[1] Здесь и далее данные в процентах указаны с учетом округления до целых.

	Реализации с бифонемным сочетанием	до[ʃ':]	Реализованы оба варианта
Иркутск	80	20	0
Ижевск	80	10	10
Казань	**100**	**0**	**0**
Воронеж	80	20	0
Псков	60	40	0
Саратов	90	10	0
Н. Новгород	**100**	**0**	**0**
В. Новгород	50	40	10
Архангельск	80	20	0

3.2. Варианты произношения слова «уезжаем»

Петербургскому варианту произносительной нормы соответствует произношение твердого согласного ([з:]) на месте орфографического сочетания «зж» внутри корня, московскому варианту — произношение мягкого [з':]. В орфоэпическом словаре под редакцией Р. И. Аванесова [9] вариант без смягчения оценивается как допустимый, согласно словарю М. Л. Каленчук, Р. Ф. Касаткиной [7] оба варианта являются равноправными. Соотношение предпочтений составило приблизительно 50/50%: *уе[з:]аем* — 56%, *уе[з':]аем* — 42%; в речи 2% информантов были реализованы оба варианта (оба варианта в речи одного и того же информанта).

Оба варианта реализаций (*уе[з:]аем* и *уе[з':]аем*) отмечены во всех регионах, однако соотношение вариантов в различных регионах варьируется: например, в Воркуте, Оренбурге, Ижевске и Н. Новгороде преобладает вариант произношения с мягким [з':], при этом преобладание, как правило, не столь сильное, как в регионах, где преобладает произношение слова «уезжаем» с твердым [з:] (Псков, Воронеж).

Таблица 4. Встречаемость вариантов произношения слова «уезжаем» в регионах в %

	уе[з:]аем	уе[з':]аем
Воркута	35	65
Оренбург	43	57
Барнаул	55	45
Иркутск	63	37
Ижевск	40	60
Казань	60	40
Воронеж	80	20
Псков	75	25
Саратов	60	40

	уе[з:]аем	уе[з':]аем
Н. Новгород	45	55
В. Новгород	65	35
Архангельск	60	40

3.3. Варианты произношения слова «отсюда»

Существуют разные точки зрения: петербургские лингвисты (например, Вербицкая Л. А.) придерживаются мнения о том, что вариант произношения слова «отсюда» с мягким [s'] (*от[s'u]да*) соответствует московскому варианту произносительной нормы, а вариант произношения с твердым [s] (*от[su]да*) — петербургскому; московские лингвисты [5, 7] основным считают вариант с мягким согласным нормативным, вариант с твердым согласным — допустимым разговорным. Вариант с твердым [s] реализовался в целом в речи лишь 10% информантов. Во всех регионах, кроме Архангельска, В. Новгорода и Барнаула небольшим количеством информантов был реализован твердый вариант:

Таблица 5. Встречаемость вариантов произношения слова «отсюда» в регионах в %

	от[s'u]да	от[su]да
Воркута	90	10
Оренбург	90	10
Барнаул	**100**	**0**
Иркутск	95	5
Ижевск	80	20
Казань	80	20
Воронеж	80	20
Псков	90	10
Саратов	95	5
Н. Новгород	85	15
В. Новгород	**100**	**0**
Архангельск	**100**	**0**

Таким образом, можно сделать вывод о том, что жители Архангельска, В. Новгорода и Барнаула с наименьшей вероятностью реализуют вариант *от[су]да*.

3.4. Ассимилятивное смягчение согласных

С точки зрения наличия или отсутствия регрессивной ассимиляции согласных по мягкости было проанализировано произношение слов «твердый»,

«дверь», «две», «свет», «свежий». В данном случае старшая языковая норма[2] допускает вариант с ассимиляцией (*[t'v']ердый, [d'v']ерь, [d'v']е, [s'v']ет, [s'v']ежий*), младшая норма — без ассимиляции (*[tv']ердый, [dv']ерь, [dv']е, [sv']ет, [sv']ежий*). Таким образом, оба варианта являются допустимыми.

Кроме того, имеет значение употребительность слова и его стилистическая окраска. «Чем частотнее в среднем слова с данным сочетанием, тем устойчивее данное сочетание» [10]. Таким образом, в более частотных словах замена первого мягкого согласного на твердый происходит раньше. К таким словам, например, относится слово «две». Аванесов обращает внимание на зависимость произношения от стилистической окраски слова. Смягчение согласного перед мягким согласным также «зависит от того, к какому стилю речи относится то или другое слово: в словах обиходно-бытового характера смягчение проводится плотнее, чем в словах книжного, в особенности иноязычного происхождения, где оно может отсутствовать» [1]. Например, слова «свежий», «свет» чаще произносятся с первым твёрдым согласным. Но допустимо и «старшее» *[s'v']ежий, [s'v']ет*.

Как указывалось выше, авторы орфоэпического словаря под редакцией Р. И. Аванесова [9] считают оба варианта произношения равноправными. В словаре М. Л. Каленчук, Р. Ф. Касаткиной [7] вариант с ассимилятивным смягчением оценивается как допустимый устаревший.

Таблица 6. Статистика произношения вариантов с наличием / отсутствием ассимиляции в %

Младшая норма		**Старшая норма**	
[dv']е	95	[d'v']е	5
[dv']ерь	96	[d'v']ерь	4
[tv']ёрдый	97	[t'v']ёрдый	3
[sv']ет	75	[s'v']ет	25
[sv']ежий	75	[s'v']ежий	25
*В речи 100% информантов из Казани и Ижевска реализовались варианты произношения слов «две», «дверь», «твердый», соответствующие младшей норме			

В произношении слов «две», «дверь», «твердый» практически полностью произошел переход к варианту младшей нормы, без ассимиляции. Тем не менее, произношение с ассимиляцией встретилось некоторое количество раз в речи жителей всех регионов за исключением регионов русско-национального двуязычия — Казани и Ижевска: все информанты из Казани и Ижевска реализовали варианты согласно младшей норме (*[dv']е, [dv']ерь, [tv']ердый*). Вряд ли в данном случае абсолютное предпочтение младшей нормы можно однозначно соотнести с двуязычным характером регионов, однако сам по себе факт примечателен.

[2] Существует понятие старшей и младшей нормы. Новое произношение (младшая норма) постепенно вытесняет старое (старшая норма), но на каком-то этапе они сосуществуют, правда, преимущественно в речи разных людей.

Менее стабильно произношение сочетания [sv'] в словах «свет», «свежий»: 25 % информантов по-прежнему произносят сочетание согласных с ассимилятивным смягчением первого элемента.

3.5. Произношение слова «судоверфь»

Регионально обусловленным также может быть произношение нечастотных слов. Одно из таких слов — слово «судоверфь» — информантам было необходимо произнести при чтении текста. Данное слово часто реализуется в речи информантов с регрессивной ассимиляцией по мягкости (*судове[r'f']*), что не соответствует норме современного русского языка, которая не допускает смягчения [r] в позиции перед мягким шумным согласным.

Таблица 6. Встречаемость вариантов произношения слова «судоверфь» в регионах в %

	судове[rf']	**судове[r'f']**
Воркута	60	40
Оренбург	65	35
Барнаул	40	60
Иркутск	85	15
Ижевск	20	80
Казань	80	20
Воронеж	65	35
Псков	65	35
Саратов	60	40
Н. Новгород	85	15
В. Новгород	40	60
Архангельск	65	35

Итак, в большинстве регионов слово «судоверфь» произносится без смягчения [r]. Однако в трех из исследованных регионов преобладает вариант с мягким [r'] — в Барнауле, Ижевске и В. Новгороде. Следует отметить, что в Ижевске преобладание ненормативного варианта является очень сильным (ненормативные реализации отмечены в речи 80 % информантов-ижевчан), поэтому для данного региона этот признак можно считать диагностическим.

3.6. Произношение слов «компьютер», «прогресс», «сервис»

В рамках данного исследования было проанализировано произношение иностранных слов (заимствований). Информантам было необходимо произнести слова «компьютер», «прогресс» и «сервис», которые на современном

этапе развития языка считаются заимствованиями. Рассматривались варианты произношения этих слов с мягким и твердым согласным перед /е/ (*компью[te]р/ компью[t'e]р, прог[re]сс/ прог[r'e]сс, [se]рвис/[s'e]рвис*). Произношение твердых и мягких согласных в заимствованных словах является очень «подвижной» нормой. Вероятно, что, заимствуясь, слово некоторое время произносится с твердым согласным перед /е/. Но по мере того, как заимствованное слово осваивается языком, оно утрачивает «налет» иностранного, и твердое произношение постепенно заменяется произношением мягкого согласного (в соответствии с правилами чтения орфографической записи).

«Словарь трудностей русского произношения» [7] дает следующую оценку нормативности вариантов произношения слов «компьютер», «прогресс» и «сервис».

Таблица 7. Оценка вариантов произношения слов «компьютер», «прогресс», «сервис» с точки зрения нормативности (согласно словарю М. Л. Каленчук, Р. Ф. Касаткиной [7])

Варианты произношения	Оценка с точки зрения нормативности
компью[te]р	нормативный
компью[t'e]р	ненормативный
прог[re]сс	нормативный
прог[r'e]сс	нормативный
[se]рвис	нормативный
[s'e]рвис	допустимый, менее предпочтительный

Как видно, оценка вариантов произношения данных слов с точки зрения нормативности авторами «Словаря трудностей русского произношения» [7] несколько расходится с оценкой их авторами словаря под редакцией Аванесова [9].

Несмотря на то, что в трех случаях мы имеем дело с произносительными вариантами одной категории, предпочтения в произношении этих слов сильно отличаются: слова «компьютер» и «прогресс», как правило, произносятся с твердым согласным (варианты *компью[te]р, прог[re]сс* реализовали около 90 % информантов). Анализ произношения этих слов показал, что зависимость выбора того или иного варианта произношения не зависит от региональной принадлежности говорящего. Однако прослеживается другая зависимость: с мягким согласным эти слова произнесли в основном люди пожилого возраста (около 50–60 лет), следовательно, можно говорить о возрастной зависимости выбора варианта произношения заимствованного слова. Слово «сервис», напротив, большинством дикторов (93 %) было произнесено с мягким согласным.

4. Выводы

В результате проведенного исследования были получены данные о региональных и индивидуальных предпочтениях в произношении слов, допускающих несколько альтернативных вариантов произношения.

Сравнивая наблюдаемые произносительные предпочтения с рекомендуемой орфоэпической «нормой», можно отметить ряд расхождений.

Так, наибольший процент предпочтений «ненормативного» (согласно словарю под ред. Аванесова [9]) или «допустимого, но менее предпочтительного» (согласно «Словарю произносительных трудностей» [7]) произносительного варианта был зафиксирован для слова «сервис», которое 93% информантов (из 240 опрошенных) произнесли со смягчением начального согласного — при рекомендуемом произнесении без смягчения. Предпочтение данного варианта наблюдалось во всех обследованных регионах.

Слово «прогресс», напротив, около 90% информантов произнесли с твердым согласным, тогда как, например, в словаре под ред. Р. И. Аванесова [9] этот вариант приводится лишь в качестве допустимого.

Слово «дождь», в произношении которого традиционно принято различать петербургскую и московскую норму, большинством информантов произносилось с двухфонемным сочетанием, т. е. в соответствии с т. н. петербургской нормой — в среднем в 75% случаев, при этом ни в одном из регионов доля данного варианта не была ниже 50%.

В отношении слов с сочетаниями гетерорганных согласных со вторым мягким можно констатировать повсеместное предпочтение произносительного варианта без ассимилятивного смягчения (что соответствует т. н. младшей норме). Однако ряд слов по-прежнему может произноситься с ассимилятивным смягчением в некоторых регионах (*свет*, *свежий* — до 25%).

В целом, по результатам проведенного исследования однозначных региональных предпочтений выявлено не было — за исключением тех немногочисленных случаев, когда тот или иной вариант произнесения в определенных регионах не встречается вообще, при его относительно низкой частотности в других регионах. В большинстве случаев скорее можно говорить о более или менее выраженной тенденции к предпочтению того или иного варианта произнесения. Для более определенных заключений и тем более для получения вероятностных мер необходимо привлекать большее число информантов и более обширную статистику произнесений. Однако и приведенные в данной работе статистические данные, как представляется, могут быть полезными при проведении идентификационного сравнения фонограмм, в том числе для целей региональной диагностики говорящего.

Отметим также, что выявленные произносительные предпочтения не всегда проявляются последовательно даже в речи одного и того же информанта, что необходимо учитывать в процессе идентификационного сравнения фонограмм.

В заключение остается добавить, что на произносительные предпочтения, особенно не ярко диалектного характера, как правило, влияет целый комплекс факторов: помимо региональной принадлежности говорящего это также уровень его образования, область его профессиональной деятельности, уровень его речевой культуры, возраст. Последнее в данной работе было проиллюстрировано произнесением слова «компьютер».

5. Перспективы исследования

Продолжение работы предполагает исследование речи жителей других регионов России с целью выявления зависимости выбора варианта произношения слов, допускающих несколько вариантов произношения. Для проведения дальнейшего исследования планируется использовать имеющуюся методику.

Изучение региональных предпочтений в произношении слов, допускающих несколько вариантов произношения, на материале речи жителей других регионов России, а также с учетом других факторов, позволит получить более полную и достоверную информацию, пригодную для целей региональной диагностики говорящего и определения других характеристик информанта. В случае обнаружения существенных межрегиональных различий эти сведения, наряду с традиционно используемыми региональными лингвистическими паспортами, могут применяться для региональной диагностики говорящего. Если же такой зависимости не будет выявлено, полученные данные относительно дистрибуции и частотности того или иного произносительного варианта могут использоваться в процессе идентификационного сравнения речевых образцов.

Литература

1. *Аванесов Р. И.* Русское литературное произношение, 6 изд., М.: Просвещение, 1984.
2. *Баранникова Л. И.* Говоры территорий позднего заселения и проблема их классификации // Вопросы языкознания, Москва, 1975. № 2, с. 22–31.
3. *Вербицкая, Л. А.* Давайте говорить правильно. Пособие по русскому языку, 3 изд., М.: Высш. шк., 2003.
4. *Ерофеева Т. И., Ерофеева Е. В., Грачева И. И.* Городские социолекты: Пермская городская речь: звучащая хрестоматия — Пермь, Бохум, 2000.
5. *Захарова К. Ф., Орлова В. Г.* Диалектное членение русского языка. М.: Наука, 1970.
6. *Исаев И. И.* Живы ли русские диалекты?: http://www.polit.ru/article/2012/03/21/isaev// (Опубликовано: 21.03.2012, 18:34)
7. *Каленчук М Л., Касаткина Р. Ф.* Словарь трудностей русского произношения, М.: Русский язык, 2001.
8. *Касаткин Л. Л.* Факторы, определяющие течение фонетического процесса — изменения С'С'>СС' в современном русском языке // Рус. яз. в науч. освещении, Москва, 2003. № 2, с. 144–172.
9. *Орфоэпический* словарь русского языка: произношение, ударение, грамматические формы / С. Н. Борунова, В. Л. Воронцова, Н. А. Еськова; Под ред. Р. И. Аванесова. Под. ред. Р. И. Аванесова. 8-е изд., М.: Русский язык, 2000.
10. *Панов М. В.* Глава пятая: Усложнение консонантной системы // Фонетика современного русского литературного языка. Народные говоры. Русский язык и советское общество: Социолого-лингвистическое исследование, Москва, 1968. с. 57–80.

ГЕНЕРАЦИЯ ШАБЛОНОВ ОЦЕНОЧНЫХ ВЫРАЖЕНИЙ НА ОСНОВЕ НЕРАЗМЕЧЕННОГО ТЕКСТА

Кравченко А. Н. (etercian@gmail.com)

Высшая Школа Экономики (ГУ), Москва, Россия

Одной из ключевых задач при извлечении мнений является отделение оценочных конструкций от фактической информации. Препятствием для решения этой задачи является высокая вариативность оценочной лексики для различных предметных областей, не позволяющая сформулировать достаточное количество универсальных закономерностей, выполняющихся для каждой области.

В данной работе предлагается алгоритм выделения шаблонов оценочных конструкций для заданной предметной области, решающий эту проблему. Алгоритм основан на методе синтаксических шаблонов и включает в себя автоматические разметку обучающего корпуса и фильтрацию сформированных шаблонов, позволяющие свести человеческое участие к минимуму.

Работа метода была проанализирована для трех различных предметных областей.

Данное исследование проводилось при финансовой поддержке Правительства Российской Федерации (Минобрнауки России) в рамках договора № 13.G25.31.0096 о «Создании высокотехнологичного производства кросс-платформенных систем обработки неструктурированной информации на основе свободного программного обеспечения для повышения эффективности управления инновационной деятельностью предприятия в современной России».[1]

Ключевые слова: извлечение мнений, определение тональности, лексические шаблоны, тональность

1. Задача отделения эмоциональной информации от фактической

В последнее время получили широкое распространение задачи оценки тональности текста и извлечения мнений (opinion mining, sentiment analysis), фокусирующихся на выделении в тексте эмоций, а не фактов. Это связано в первую очередь с возросшей популярностью блогов и социальных сетей и увеличением объемов пользовательского контента. Мнения — источник ценной информации как для социологов, маркетологов и журналистов, так и для самих пользователей, при этом поиск подобной информации часто затруднен.

[1] Доклад публикуется в сокращенном варианте. Полная версия доступна на сайте конференции «Диалог».

Полезной может оказаться небольшая часть длинного сообщения или одно сообщение из нескольких страниц темы форума.

Наибольшую сложность в решении этой задачи представляет отделение эмоциональной информации от фактической и выделение субъективных конструкций. Эффективным методом являются лексические шаблоны, но на практике возникает препятствие, связанное с высокой вариативностью оценочной лексики для различных предметных областей. Например, выраженная одной и той же фразой оценка может быть положительной в одном случае и отрицательной в другом («непредсказуемый сюжет фильма» и «непредсказуемое поведение программы»), или не являться оценкой вовсе. Таким образом, составленный для одной предметной области набор шаблонов нельзя использовать даже для смежных областей, что делает ручное составление шаблонов неоправданным. Другим серьезным препятствием является недостаток обучающих данных.

В данной работе предлагается решение этой проблемы с помощью алгоритма извлечения шаблонов оценочных выражений для заданной предметной области. Алгоритм основывается на выявленных при изучении тонально окрашенных текстов закономерностях, таких как схожесть синтаксической структуры оценочных текстов и повторяемость типичных конструкций. Благодаря использованию этих закономерностей, работа алгоритма не требует предварительной разметки корпуса, составления словаря предметной области и сводит к минимальной необходимость фильтрации результата.

Полученные шаблоны могут использоваться для получения отзывов в интернет-магазинах для заданной категории продуктов или для выявления тенденций в блогах или новостных обзорах.

2. Метод лексических шаблонов

Метод, предлагаемый в данной работе ориентирован на извлечение оценочных фрагментов из текстов произвольного формата. Это может быть полезно, например, для обработки сообщений в блогах, где много «лишней» информации, или в тех случаях, когда недостаточно узнать только общую оценку объекта и нужен подробный анализ (автор отзыва может быть в целом доволен объектом, но критиковать отдельные его свойства, или наоборот).

Использовать лексические шаблоны в данном случает позволяет то, что синтаксическая структура оценочных конструкций, в отличие от лексики, практически не зависит от предметной области.

Предлагаемый метод состоит из трех шагов:

Шаг 1: Производится разметка частей речи и разбивка предложений на сегменты.

Определение: Сегментом будем называть предложение или фрагмент предложения, не содержащие противопоставлений.

Отсутствие противопоставлений внутри сегмента в большинстве случаев позволяет однозначно оценить его эмоциональную окраску.

На первом шаге в качестве сегментов будем выделять простые предложения, сложные предложения, не содержащие противопоставлений и части сложных предложений, разделенные союзами «но», «зато», «несмотря», «хотя», «однако».

Каждый сегмент представляется как последовательность, являющаяся «заготовкой» для шаблона. Например:

«Это плохие наушники»

Превращается в последовательность:

{это, **PRN}{**плохие, **ADJ}{**наушники, **NN}**

Шаг 2: Используются n-граммы чтобы разбить сегменты на более короткие. Триграмм обычно оказывается достаточно. Если не использовать n-граммы, длинные сегменты приводят к порождению неподходящих, редко встречающихся или просто некорректных шаблонов.

Шаг 3: Все слова, обозначающие свойства объекта, заменяются меткой свойства **$feature**. Эта замена необходима, т. к. свойства разных продуктов разные, и замена гарантирует, что мы сможем найти общий для всех шаблон.

Свойствами будем считать все возможные компоненты (например, для фотоаппарата это будут объектив, батарея, видоискатель) и атрибуты (качество картинки, вес, размер фотоаппарата) объекта, или название самого объекта.

После замены указанная выше последовательность превращается в:

{это, **PRN}{**плохие, **ADJ}{$feature**, **NN}**

Результат сохраняется в базе, к нему также может применяться стемминг. Такая последовательность называется ***лексическим шаблоном***.

После этого можно использовать полученные последовательности для извлечения данных. Совпадения с лексическими шаблонами ищуются в каждом сегменте отзыва. Слово, совпадающее с меткой **$feature** считается свойством.

3. Классификация оценочных конструкций

Тонально окрашенными будем называть такие элементы текста (синтагмы, фразы), которые несут в себе оценочную семантику.

Важными являются понятия ***пропозиции,*** в том числе ***логической*** и ***событийной***, *и* ***семантических ролей,*** используемые в формальной семантике. ***Пропозицией*** называется обозначенное в речи действительное или возможное положение дел, семантическое ядро высказывания. Основой пропозиции является предикатное выражение, которое состоит из предиката,

выражающего действия, состояния, свойства или отношения, и набора его аргументов, которые представляют собой знаки вещей, то есть имена, чьи свойства обозначаются.

Предикат ***событийной пропозиции*** связан со сферой бытия, движения, деятельности (физической или социальной); **логической пропозиции** — с отражением отношений, устанавливаемых в процессе мыслительной деятельности, логических рассуждений (отношения идентификации, тождества и т. п.).

Семантической ролью имени называется часть семантики предиката, отражающая общие свойства аргумента предиката — участника называемой предиктом ситуации. Описание в терминах семантических ролей отражает сходства моделей управления различных предикатных слов. Это понятие было предложено Чарльзом Филлмором, вначале использовавшим термин «глубинный падеж».

Теория семантических ролей слишком объемна для того, чтобы приводить ее здесь полностью. За подробным изложением этой теории можно обратиться к книге [Гриднева, 2009] или [Сусов, 2006].

В работе Ермакова и Киселева [Ермаков А. Е., Киселев С.Л, 2005] была предложена следующая классификация оценочных конструкций:

Тип 1. Явная тональная характеристика

Объект или инициированное им событие наделяется признаком, имеющим оценочную семантику.

Типовые пропозиции, которыми выражаются ситуации этого класса, курсивом выделены тонально окрашенные участники:

1. Логическая пропозиция полная: наушники *дрянь*, Сидоров *плохой руководитель*. Тональность выражается именной группой, образуемой существительным.

2. Логическая пропозиция свернутая с существительным: *гениальный авантюрист* Петров; *кристальная чистота* звука. Тональность выражается именной группой, образуемой существительным.

3. Логическая пропозиция свернутая с прилагательным: *запоминающийся* дизайн, *качественное* изображение, руководитель *нерешителен*. Тональность выражается прилагательным.

4. Свернутая логическая пропозиция в составе событийной, отражающая оценку события, в котором объект оценки выступает в роли протагониста: телефон *быстро сломался*, Иванов *бездумно согласился,* президент *принял авантюрное решение*. Тональность может выражаться наречием при глаголе, прилагательным при событийном существительном, самим глаголом.

Тип 2. Прямая эмоционально-коннотативная характеристика

Класс эмоционально «заряженных» ситуаций, отражающих отношение целевого объекта к тонально окрашенным сущностям, их оценку целевым объектом, или наоборот отношение этих сущностей к объекту, оценку объекта

ими. Выражается событийными пропозициями: президент борется с преступностью, народ выносит осуждение власти.

Тип 3. Ассоциированный эмоциональный коннотат

Класс эмоционально заряженных ситуаций, фигурирующих в одном предложении с целевым объектом, но не связанных с ним напрямую (в ряде случаев эту связь просто не удается идентифицировать средствами автоматического анализа текста). Выражаются событийными пропозициями: обнищание пенсионеров, купил новую пару, пристраститься к пиву.

Анализ типов 2 и 3 методом шаблонов сильно затруднен, так как их обработка требует глубокого синтаксического и семантического анализа. В дальнейшем в данной работе будут рассматриваться только оценки первого типа.

4. Определение границ мнения и преобразование полученных оценочных конструкций в шаблоны

На основании приведенной в предыдущем разделе классификации были разработаны синтаксические шаблоны, используемые для выделения оценочных конструкций из текста.

Приведем примеры для каждого типа оценки:

Пример:

(1) *Логическая пропозиция полная:*
наушники — дрянь — {Nn-Nom, Nn-Nom}
Иванов — плохой руководитель — {Prop-Nom, Adj-Nom, Nn-Nom}

Пример:

(2) *Логическая пропозиция свернутая с существительным:*
кристальная чистота звука — {Nn-Nom, Adj-Nom, Nn-Gen}

Пример:

(3) *Логическая пропозиция свернутая с прилагательным:*
качественное изображение — {Adj-Nom, Nn-Nom}
руководитель нерешителен {Nn-Nom, Adj-Brf}

Пример:

(4) *Свернутая логическая пропозиция в составе событийной:*
телефон быстро сломался — {Nn-Nom, Adv, Verb}
Иванов бездумно согласился {Prop-Nom, Adv, Verb}

При обучении алгоритма находятся и учитываются все совпадения синтаксических шаблонов с сегментами, определенными как оценочные, затем

полученный сегмент преобразуется в лексический шаблон. Как правило, это происходит с помощью замены существительного или глагола на метку атрибута.

Заметим, что применение этого метода требует предварительного морфологического анализа текста.

5. Подготовка обучающего корпуса

Для обучения разработанного алгоритма необходим корпус, к котором каждое предложение размечено как оценочное или безоценочное. Определения тональной направленности документа в целом недостаточно, так как то, что текст в целом является оценочным, не означает, что оценочным является каждое его предложение. Субъективные тексты обычно содержат какое-то количество фактической информации.

Одним из основных препятствий для создания алгоритмов, подобных описанному, является недостаток обучающих данных. Для обучения алгоритма, классифицирующего тексты, легко составить коллекцию, используя, например, отзывы на соответствующих сайтах, многие из них помимо отзыва включают в себя оценочные баллы.

Гораздо тяжелее составить коллекцию отдельных предложений, которые могут быть однозначно определены как объективные или субъективные. Большая часть работ, посвященных алгоритмам, работающим на уровне предложений, использовали вручную размеченный корпус. Это требует больших временных и трудозатрат, поэтому объем подобных корпусов сравнительно мал.

Elen Riloff [Riloff&Wiebe, 2003] был предложен метод автоматизированной разметки, позволяющий создать корпус гораздо большего объема, чем любой из размеченных вручную, так как количество неразмеченных оценочных текстов, например, в Интернете огромно. Он предполагает использование контекстно-независимых оценочных слов для поиска оценочных предложений. Если предложение содержит такое слово, оно считается оценочным, в противном случае предложение считается нейтральным.

Заметим, что в предложении могут одновременно содержаться как контекстно-зависимые, так и контекстно-независимые оценочные слова, что позволяет, распознав оценочное предложение по контекстно-независимому слову, выделять шаблоны, специфичные для заданной предметной области.

6. Фильтрация полученных шаблонов

Отбор работающих шаблонов обычно предполагает участие человека, данный алгоритм позволяет его минимизировать при наличии достаточно большого обучающего корпуса.

Метод основывается на том наблюдении, что оценочные конструкции, как правило, встречаются повторно в других оценочных текстах. Если конструкция

была выделена как оценочная ошибочно, скорее всего, она не встретится или будет встречаться реже настоящих оценочных.

Для каждого шаблона по следующей формуле вычисляется условная вероятность того, что предложение, содержащее этот шаблон, будет оценочным:

$$P(subjective/pattern_i) = subjfrec(pattern_i)/frec(pattern_i),$$

где subjfreq(pattern) — количество раз, которые шаблон встретился в оценочных сегментах, freq(pattern) — количество всех использований шаблона.

Если вероятность выше порогового значения, шаблон принимается, если ниже отбрасывается. Пороговое значение выбирается экспериментально.

7. Схема работы алгоритма

Таким образом, работа алгоритма состоит из 6 этапов:

1. разметка частей речи
2. обобщение морфологических категорий
3. классификация сегментов
4. выделение оценочных фрагментов
5. преобразование текстовых фрагментов в шаблоны
6. фильтрация шаблонов

Для разметки частей речи документ пропускается через морфологический парсер с параметрами сохранения пунктуации и выдачи грамматических категорий. После этого каждая лексема представляется в виде структуры <лексема, базовая форма, тип лексемы>, а сам документ разбивается на сегменты. Окончанием сегмента считается точка, многоточие, точка с запятой, восклицательный или вопросительный знак, или запятая за которой следует слово «но», «зато», «хотя», «однако».

Классификация лексем основана на диссертации И. Ножова [Ножов И., 2003].

Каждый сегмент классифицируется в соответствии с описанным в разделе 3 методом. Если сегмент содержит оценочное слово, он считается оценочным, если нет нейтральным. Список оценочных слов составлен вручную с помощью словарей синонимов и содержит около 500 слов, все они являются контекстно-независимыми.

После этого к полученным сегментам применяются синтаксические шаблоны, представляющие из себя последовательности типов лексем, сопоставляемых с типами лексем в тексте. Находятся и учитываются все совпадения. Экспериментальная оценка показала, что не имеет смысла использовать шаблоны, содержащие более четырех элементов.

Затем из каждого полученного сегмента в соответствии с типом синтаксического шаблона выделяется лексический шаблон оценочного выражения.

Далее по приведенной в предыдущей главе формуле вычисляется вероятность того, что шаблон является оценочным. Каждое слово шаблона

сопоставляется с базовой формой встреченного в тексте слова, сохраняется согласование падежей внутри шаблона. Находятся все совпадения.

$$Pr(subjective|patter_i) = subjfrec(pattern_i)/freq(pattern_i),$$

где subjfreq(pattern) количество совпадений шаблона с оценочными сегментами, freq(pattern) количество всех совпадений шаблона с текстом.

Если вероятность не превышает заданное пороговое значение, шаблон отбрасывается. Эксперименты показали, что для более формальных областей, например, бытовой техники оптимальное пороговое значение может превышать 0.9, для областей, предполагающих менее четкую терминологию (таких как, например, отзывы о фильмах) снижаться до 0.6.

Заметим, что для корректной работы фильтра корпус должен быть достаточно большим, не менее 100 000 предложений.

Пример работы алгоритма:

Исходный текст (комментарий в интернет-магазине, пунктуация сохранена):

«Купил вчера эти наушники. Звук — сказка хирургическая точность. Сочные обволакивающие басы, отличные верха и громкость.

Один недостаток — конструкция немного хрупковатая.»

Этапы 1–3: Морфологическая разметка и сегментирование текста (жирным шрифтом выделены сегменты, отмеченные как оценочные):

[< Купил купить Verb-Fin > < вчера вчера Adv > < эти этот Det-Nom > < наушники наушник Nn-Nom >]

[< Звук звук Nn-Nom > < сказка сказка Nn-Nom > < хирургическая хирургический Adj-Nom > < точность точность Nn-Nom >]

[< Сочные сочный Adj-Nom > < обволакивающие обволакивать Adj-Nom > < басы бас Nn-Nom >

< отличные отличный Adj-Nom > < верха верх Nn-Nom >

< и и Conj > < громкость громкость Nn-Nom >]

[< Один один Det-Nom > < недостаток недостаток Nn-Nom >]

[< конструкция конструкция Nn-Nom > < немного немного Adv >

< хрупковатая хрупковатый Adj-Nom >]

Этапы 4–5: Выделение оценочных фрагментов с помощью синтаксических шаблонов, преобразование их в лексические шаблоны:

[Adj-Nom, Nn-Nom, Conj, Nn-Nom] -> отличные верха и громкость ->
отличные $feature

[Adj-Nom, Adj-Nom, Nn-Nom] -> сочные обволакивающие басы ->
сочные обволакивающие $feature

[Nn-Nom, Nn-Nom] -> звук сказка -> **$feature сказка**

[Nn-Nom, Adj-Nom] -> сказка хирургическая -> **$feature хирургическая**

[Adj-Nom, Nт-Nom] -> хирургическая точность ->
хирургическая $feature
[Nn-Nom, Adv, Adj-Nom] -> конструкция немного хрупковатая ->
$feature немного хрупковатая

Этап 6: Фильтрация шаблонов
оставлены:
отличные $feature
сочные обволакивающие $feature
$feature немного хрупковатая

отфильтрованы:
$feature хирургическая
хирургическая $feature

8. Оценка результата

В настоящее время методов объективного тестирования систем тональной разметки текстов еще не разработано. Поэтому применяемый нами метод тестирования основывается на субъективных оценках небольших текстовых подборок экспертом. В качестве тестовой коллекции был собран корпус объемом приблизительно 300 000 предложений, состоящий из трёх частей — отзывы о фотокамерах, отзывы о наушниках и рецензии зрителей на фильмы.

Эксперт получает тональную разметку текстов при помощи системы и затем оценивает, насколько он в каждом конкретном случае согласен или не согласен с результатом. В случае несогласия эксперт отмечает, что нетональная конструкция была размечена как тональная. Затем на основании этих оценок вычисляется точность тональной разметки по формуле:

$$P= N_subj/N_all,$$

где N_subj количество верных шаблонов N_all количество всех выделенных шаблонов.

При оценке мы не учитываем полноту, т. к. из-за самой конструкции алгоритма она является невысокой (прежде всего из-за использования автоматической фильтрации), что может быть компенсировано размером обучающей коллекции благодаря автоматической разметке.

Примеры извлеченных конструкций :

(6) *отзывы о фотоаппаратах:*
$feature недостаточно светосильный
хороший резкий $feature
прочный удобный $feature

удобная быстрая $feature
дополнительный монохромный $feature
яркий большой $feature
неравномерное искусственное $feature
избыточное число $feature
низкий уровень $feature

(7) ***рецензии на кинофильмы:***
слишком поверхностные $feature
очень жесткий $feature
неприятный $feature
актуальный $feature
весьма странный $feature
страшный $feature
довольно тяжелый $feature
$feature немного нудный
больное воображение $feature
великолепная игра $feature
$feature гениальный
смешное такое $feature

(8) ***отзывы о наушниках:***
кристально чистый $feature
более обьемный $feature
$feature немного хрупковатая
$feature вполне приличный
$feature весьма ничего
$feature просто супер
$feature ничего особенного
профессиональная звуковая $feature
пластиковый прочный $feature

Точность выдачи алгоритма оказалась зависящей от предметной области.

Для областей, подразумевающих большое количество объективных критериев и относительно невысокую вариативность лексики, таких как отзывы о наушниках и фотокамерах, алгоритм показал очень хорошие результаты до 52% до фильтрации и 80% после фильтрации с пороговым значением вероятности сохранения шаблона $\theta=0.9$.

Для рецензий на фильмы области с большей вариативностью лексики и небольшим количеством объективных критериев точность была гораздо ниже 29% до фильтрации и 64,3% после фильтрации с пороговым значением вероятности сохранения шаблона $\theta=0.6$.

Точность можно повысить, увеличивая для алгоритма фильтрации пороговое значение вероятности сохранения шаблона, но это сильно снижает полноту, особенно для плохо формализованных областей.

Также можно выделить два класса ошибок, возникающих при определении субъективности:

1. Ошибки работы модуля морфологической разметки текста.
2. Ошибки, связанные со сходством фактических конструкций с оценочными.

9. Заключение

Как показывают результаты, предложенный метод достигает достаточно высокой (80%) точности на текстах определенной тематики и может быть в дальнейшем использован в качестве компонента системы извлечения мнений. Точность так же может быть повышена с помощью улучшения качества разметки обучающего корпуса и использованиея более глубокого синтаксического анализа, что является направлением дальнейшей работы.

References

1. *Carenini G., Pauls A.* (2006), Multi-Document Summarization of Evaluative Text in Proceedings of EACL 2006, Trento, Italy, 2006, pp. 305–312.
2. *Ermakov A. E., Kiselev S. L.* Linguistic Model For Computational Sentiment Analysis of Media [Lingvisticheskaya model dlja komputernogo analiza tonalnosti publikatsij v SMI]. Kompjuternaja Lingvistika i Intellektualnye Tehnologii: Trudy Mezhdunarodnoj Konferentsii Dialog 2005 [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialog 2005, Moskva, 2005.
3. *Gridneva N. N.* (2009) Osnovy semantiki sintaksisa [Introduction to Syntax and Semantics]. St.Petersburg, SpGUEF.
4. *Hatzivassiloglou V., McKeown K.* (1997), Predicting the semantic orientation of adjectives in Proceedings of ACL/EACL 1997, Madrid, Spain, Complutense University of Madrid, 2007, pp. 174–181.
5. *Hu M. and Liu B.* (2004), Mining and Summarizing Customer Reviews in Proceedings of KDD-2004, Seattle, WA, 2004, pp. 168–177,
6. *Hu M. and Liu B.* (2004), Mining Opinion features in Customer Reviews in Proceedings of AAAI'04, Boston, Massachusetts, USA: AAAI Press, 2004, pp. 755–760.
7. *Kobayashi N., Inui K., Tateishi K., Fukushima T.* (2004), Collecting Evaluative Expressions for Opinion Extraction in Proceedings of IJCNLP-2004, Berlin, Germany: Springer, 2004, pp. 596–605.
8. *Liu, B.* (2007), Web Data Mining, Springer, Berlin.
9. *Liu, B.* (2010), Sentiment Analysis and Subjectivity in Handbook of Natural Language Processing, Second Edition, Chapman and Hall/CRC 2010, NY, USA, pp. 257–282.
10. *Nozhov I.* (2003) Morphologic and Syntactic Text Processing (Models and Computations), available at http://www.aot.ru/docs/Nozhov/msot.pdf

11. *Popescu A. M., Etzioni O.* (2005), Product features and Opinions from Reviews in Proceedings of HLT-EMNLP 2005. Vancouver, Canada: ACL, 2005, pp. 339–346.
12. *Riloff E, Wiebe J.* (2003) Learning Extraction Patterns for Subjective Expressions, in Proceedings of EMNLP-03, Sapporo, Japan, 2003, pp. 97–104.
13. *Sibiriakov A.* Extracting opinions about products from blogs and forums according to sentiment [Izvlecheniye mneniy o tovarah iz forumov i blogov s uchetom tonalnosti], available at http://elar.usu.ru/bitstream/1234.56789/2064/1/RuSSIR_2008_07.pdf
14. *Susov I. P.* (2006) Vvedeniye v Yazykoznaniye [Introduction to Lingustics]. Moscow, Vostok-Zapad.
15. *Turney P.* (2003), Inference of Semantic Orientation from Association, available at http://cogprints.org/3164/01/turney-littman-acm.pdf
16. *Turney P.* (2002), Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews in Proceedings of ACL 2002, Philadelphia, PA, U.S.A., 2002, pp. 417–424.

THE GENERAL CORPUS OF THE MODERN MONGOLIAN LANGUAGE AND ITS STRUCTURAL-PROBABILISTIC MODEL[1]

Krylov S. A.

Institute of Oriental Studies, Russian Academy of Sciences, Moscow, Russia

The paper describes the General Corpus of the Modern Mongolian language (GCML), which contains 966 texts, 1 155 583 words. We also report a morphological analyzer for Modern Mongolian language (MML), a grammatical dictionary for 63 071 lexemes, a general table of morphological homonymy. The processor analyzes effectively 97 % of textual word forms which correspond to 76 % word forms from the inputs of the concordance to the GCML. MML can be described in its quantitative aspect, according to a structural-probabilistic model (SPM) of MML. SPM contains frequency dictionaries (FDs) of MML of different types: FDs of word forms, lexemes, grammatemes, root morphemes and allomorphemes, affixal morphemes and allomorphemes, flexionemes, grammemes.

SPM allows describing behavior of various language units in the written text from the quantitative point of view: their frequency, distribution in texts, compatibility with other units etc. It is possible to transform the usual structural model into a SPM, which is based on statistical analysis of texts (in this model units of language are considered as possessing "their weight", the language oppositions and relations are being measured).

The paper reports the top lists of some FDs: i. e. ranging FD of word forms (top-list of the upper 44 word forms having frequencies higher than 1700 ipm), ranging FD of lexemes (top-list of the upper 44 lexemes having frequencies higher than 2050 ipm) and ranging FD of grammatemes (top-list of the upper 44 grammatemes having frequencies higher than 2909 ipm).

Key words: corpus linguistics, modern Mongolian language, frequency dictionaries, quantitative approach in linguistics

[1] The present research has been done under the financial support of the Program of the Presidium of the Russian Academy of Sciences on corpus linguistics on 2011. (direction 4 — "Creating and development of corpus resources on the languages of the world". The corpus has been created together with doctor of philology G. Ts. Pyurbeev, candidate of philology Natalya S. Yakhontova and candidate of philology Maria P. Petrova, to whom the author of the present article expresses a cordial gratitude.

1. About the general corpus of Mongolian

The initial version of the corpus of the modern Mongolian language reported here includes the following text genres:

1) fiction texts of the XX century: novels and stories; sketches;
2) poetry of the XX century;
3) «The Secret Legend of Mongols» epos translated into the modern language;
4) selection of newspaper articles from the newspaper «Dajaar Mongol».

The corpus contains 966 texts (1 155 583 words length).

A morphological analyzer, a dictionary for 63 071 lexemes, a table of homonyms have been created. The corpus has been lemmatized and glossed (in the spirit of the Leipzig glossing rules[2]).

The morphological analyzer at present works within the StarLing environment[3]. At present the work is on the experimental stage, 97 % of text word forms (which correspond to 76 % of the word forms which are inputs in the concordance of word forms) can be effectively analyzed.

Today the total analysed (morphologically marked) graphic word forms (allolexes) in the corpus is 1 103 233. The total of graphic word forms (lexes) in the corpus is 1 155 583, i. e. 97 %.

In total the amount of different allolexemes in the vocabulary of the concordance to the corpus (and in vocabulary of allolexemes) is 89 190. The share of the recognized allolexemes is 67 531, i. e. 76 %.

Overall effectiveness of the morphological analyzer can be visually presented in table 0.

Table 0.

	In the corpus:	part in %	In the vocabulary:	part in %
total word forms:	1 155 583	100 %	89 190	100 %
analysed:	1 123 156	97 %	79 137	89 %
analysed lexically:	1 104 911	96 %	68 212	76 %
analysed grammatically:	1 121 478	97 %	78 456	88 %
analysed both lexically, and grammatically:	1 103 233	95 %	67 531	76 %

[2] See Lehmann 1982; Croft 2003; Kassevich 2011: 214–221; also www.eva.mpg.de/lingua/resources/glossing-rules.php. A case study of the use of such notation in Mongolic studies (concerning a corpus with a resolved homonymy) see, for example, Baranova an Say 2009: 10–16, 873.

[3] The StarLing software environment was created by Sergey A. Starostin (1953–2005), and later it's been maintained by Philipp S. Krylov.

2. The Mongolian language in quantitative aspect

On the materials of the General Corpus of the Mongolian language first attempts of describing the Mongolian language in its quantitative aspect have been done. The Structural-probabilistic model of the Mongolian language includes frequency dictionaries (FD) of the Mongolian language of different types: frequency dictionaries of word forms in the ranging and in the alphabetic order (direct and inversed), FD of the bases in ranging and in alphabetic order (direct and inversed), FD of inflections in ranging and in alphabetic order (direct and inversed), FD of grammatemes in ranging and in alphabetic order (direct and inversed), and in an ideographic order, FD of lexemes in ranging and in alphabetic order (direct and inversed), FD of flexionemes in ranging and in alphabetic order (direct and inversed), FD of affixal allomorphemes and of affixal morphemes in ranging and in alphabetic order (direct and inversed), FD of root allomorphemes and of root morphemes in ranging and in alphabetic order (direct and inversed), FD of grammemes in ranging, alphabetic, and in an ideographic order.

3. The Structural-probabilistic model of Mongolian

The problem of studying of the quantitative characteristics of the MML is urgent because the majority of these characteristics till now are unknown to the scientists in the absence of a representative and at least relatively balanced corpus of MML to provide the material to apply distributive-statistical methods allowing to make professionally high-quality FDs and quantitative grammars, describing rate of units of morphology, derivatology, syntax and lexicology.

The quantitative approach allows to classify texts according to language styles and genres in frameworks of which these texts were created. As the distinctions between these styles and genres «are mainly a statistical quality»[4] thus it is possible to base the statistical stylistics of the MML describing and classifying texts of the MML on the strictly objective basis.

The quantitative approach to texts opens a way to studying MML while the segments of texts which are objects of calculations, are correlated with the units of the MML. The Linguo-statistical method allows to describe behavior of different language units (letters, morphemes, words etc.) in written text from the quantitative point of view: frequency of the use of units, their distribution in texts of a different genre, compatibility with other units etc. “At the same time the generalized quantitative information on classes of units, on language designs (e. g., the data about average length of a word or a sentence, on the frequency of the use of grammatical forms in different syntactic functions etc.) is being accumulated. Such information deepens the description of units of language”[5]. For example, simple ascertaining of existence

[4] See Shaykevich 1990: 231.

[5] See Shaykevich 1990: 231.

of forms of plural of nouns in RL and ML is insufficient for revealing of typological distinctions if one does not consider quantitative distinctions in the text functioning of the corresponding units. «This makes it possible to transform the usual structural model of language into a structural-probabilistic model, which is based on statistical analysis of texts (in this model units of language are considered as possessing "their weight", the language oppositions and relations are being measured). The structural-probabilistic model of language is more realistic, it is especially effective in diachronic and typological studies (.)»[6].

4. Frequency dictionaries FDs of Mongolian

Below we provide examples of the top lists of some FDs. Numerical indicators in column C mean relative frequency (quantity of occurrences of the given unit per one million word forms[7]), in column D — quantity of texts in which the given unit occurs, in column E – the rank of the given unit.

4.1. Frequency of word forms in Mongolian

Tab. 1 provides frequencies of word forms. Column A — a word form; column B gives the approximate English translation of the word of ML (in its main meaning).

Table 1.

A	B	C	D	E
нь	his, her, its, their *[also def. art.]*	24463.32	666	1
гэж	that *[conjunction for the object clause]*	13884.12	478	2
юм	*[marker of rheme]*	10052.32	529	3
ч	the	9882.74	540	4
л	let	6798.44	441	5
энэ	this, he, she, it, they	6628.87	463	6
тэр	that, he, she, it, they	5801.78	421	7
нэг	one *[actantial]*	5658.16	399	8
би	I *[subject]*	5429.76	446	9
хүн	man, person	5260.18	493	10
байна	is, are, am	4850.10	437	11
шиг	like, similar	4619.10	526	12
хоёр	two, both, and	4590.55	409	13
минь	my *[also a def. art.]*	4161.43	496	14

[6] See Shaykevich 1990: 231.

[7] On the use of "ipm" unit (instances per million words) see Sharoff 2002; Sharoff and Lyashevskaya 2009: 9.

A	B	C	D	E
байгаа	be *[ger. imperf.]*	3666.56	340	15
бол	as to; *[marker of topic]*	3606.86	363	16
дээр	on	3339.53	461	17
чинь	your *[also a def. art.]*	3109.39	344	18
байсан	was, were *[ger. perf.]*	2916.46	337	19
их	big, very	2907.81	403	20
дээ	let it be	2858.49	293	21
юу	what *[nom.case]*; whether	2722.66	307	22
гээд	having said; having told	2625.77	275	23
чи	you *[subject.]*	2620.58	289	24
уу	whether	2414.67	321	25
бас	too; also	2359.30	349	26
билээ	is, are, am	2338.53	303	27
байх	be	2313.44	316	28
сайхан	nice, beautiful	2217.41	429	29
байлаа	was, were	2190.59	265	30
шүү	not so?; isn't it?; doesn't it? wasn't it?; weren't it? didn't it?; etc.	2188.86	264	31
гэсэн	has told; has said	2091.10	314	32
вэ	*[special question]*	2049.57	283	33
болж	becoming	2018.42	322	34
биш	not	1965.65	306	35
та	you *[subject.]*	1965.65	259	36
гэдэг	they say	1925.85	322	37
одоо	now	1884.32	283	38
миний	my, mine	1876.54	330	39
хар	black; look!	1816.84	315	40
газар	land; country	1812.51	331	41
үгүй	no	1795.21	282	42
хүний	of a man, of a person *[gen.case]*	1760.60	389	43
болсон	become	1700.91	317	44

4.2. Frequency of lexemes in Mongolian

Tab. 2 represents frequency of lexemes. In column A lexemes are given; in column B approximate English translations of the word of ML (in its major sense) are given.

As the work was done on the corpus with unresolved homonymy sometimes the status of lexemes is ascribed not actually to lexemes, rather to disjunctive bundles of (partially) homonymic lexemes. However the information on the frequency of such units in the corpus is not less important than the information on the frequency of the

real lexemes. Anyway, for the information revealing quantitative characteristics of lexicon and grammar of the ML it is not necessary to wait, when the corpus with the unresolved homonymy will be created: it is necessary to wait too long, and anyway the data received on the basis of the corpus with the resolved homonymy will be based on too small empirical facts that will depreciate their statistical significance.

Table 2.

A	B	C	D	E
нь	his, her, its, their *[also def.art.]*	24463.32	666	1
гэж	that *[conjunction for object clause]*	13887.58	478	2
байх	be	13638.41	525	3
юм~юм(ан)	*[marker of rheme]*; thing	10408.76	531	4
ч	the	9882.74	540	5
болох	become	9672.51	528	6
явах	go	7327.06	511	7
л	let it	6845.16	443	8
энэ	this; he; she; it	6636.66	463	9
тэр	that; he; she; it	6024.99	427	10
нэг(эн)	one	5749.00	402	11
хүн	person; man	5538.77	497	12
ирэх	come	5497.24	439	13
би	I	5434.95	448	14
байн~байх	be	5364.87	449	15
хоёр	two; both; and	5197.03	427	16
шиг~шигэх	like; similar	4624.29	526	17
дээр	on	4445.20	503	18
байг~бай~байх	be	4322.35	362	19
минь	my *[also def. art.]*	4163.16	497	20
хэлэх	speak; talk	3645.79	348	21
бол	as for *[marker of topic]*	3609.46	363	22
гарах	come out of	3329.14	392	23
чинь	your *[also def.art.]*	3110.26	344	24
орох	enter	3051.43	363	25
гэх	say; tell	3019.42	328	26
их	very; big	3011.63	411	27
дээ	let it	2863.69	294	28
юу~юу(н)	what? whether	2725.26	307	29
чи	you *[sg.]*	2661.24	307	30
бодох	think; count; calculate; intend; aspire	2637.88	303	31
гээд~гээ~гэх	having told; having said	2627.50	275	32
сэтгэл	thaought; intention; wish	2464.85	405	33
сайхан	beautiful; nice	2448.41	458	34

A	B	C	D	E
уу~уух	whether; drink!	2424.18	322	35
харах	look	2386.12	343	36
бас	too; also	2365.35	350	37
билээ	is; are; am	2340.26	303	38
мэдэх	know	2265.86	316	39
үзэх	see	2157.71	355	40
өгөх	give	2127.43	307	41
гэсэх~гэх	say; tell	2101.48	314	42
вэ	*[marker for special question]*	2052.16	284	43
гэдэг~гэх	they say	2050.43	325	44

A lot of useful information can be drawn based on top-lists of language units in ranging dictionaries. Such information gives impulse to some interesting observations of typological nature. Below we attempt to present some of such observations.

The conjunction *бөгөөд* 'and' occupies the 52nd position in the FD of wordforms and the 67th position on the FD of lexemes.

Compare: the conjunction *and* in English FDs occupies the 3rd (or 4th , or 5th) position; conjunction *и* ('and') in Russian FD occupies the 1st position.

Why such difference? The difference can be explained in the following way. Mongolian prefers asyndetic constructions. The frequent asyndeton usually is being compensated through the phenomena of the so-called Altaic type of coordination which means the group inflection of nouns and the rich system of gerundial taxis constructions in the domain of verb.

2. Possessive pronouns have extraordinary high frequencies.

E.g., *нь* (his, her, its, their) occupies the 1st position in the FD of lexemes and the 1st position in the FD of word-forms. Compare: in English FDs *his* occupies the 25th (or the 23rd, or the 12th) position, *her* occupies the 42nd (or the 29th, or the 13th) position, *its* occupies the 78th (or the 77th, or the 142nd) position, *their* occupies the 36th (or the 39th, or the 61st) position. In Russian FDs *его* ('his', 'its') occupies the 41st (or the 50th) position, *её* ('her') occupies the 72nd (or 121st) position, *их* ('their') occupies the 86th (or 134th) position.

минь ('my') occupies the 14th position in the FD of word-forms and the 20th position in the FD of lexemes. Compare: in English FDs *my* occupies the 44th (or the 34th, or the 24th) position. In Russian FDs *мой* ('my') occupies the 60th (or the 69th) position

чинь ('your') occupies the 18th position in the FD of word-forms and the 24th position in the FD of lexemes. Compare: in English FDs *your* occupies the 69th (or the 64th, or the 62nd) position. In Russian FDs *твой* ('your') occupies the 266th (or 579th) position

Why such difference? The difference can be explained in the following way. Mongolian has a grammatical category of possessiveness. It can be expressed synthetically or analytically. Analytic means of expression of the category are the encliticized possessive pronouns. In fact the encliticized possessive pronouns play a kind of role which is fulfilled by articles in the European (e. g., Romance and Germanic) languages with articles. If one compares the European possessive pronouns with the Russian ones,

one can see that the European possessive pronouns are used more frequently than the Russian ones (in Russian-English translations such facts can be seen especially clear: compare such translational equivalents as *жена* ó *my wife* (*your wife*, *his wife*), *мать* ó *my mother*(*your mother*, *his mother*, *her mother, our mother, their mother*), *муж* ó *my husband*(*your husband, her husband*) , *отец* ó *my father*(*your father, his father, her father, our father, their father*), *нос* ó *my nose*(*your nose, his nose, her nose*), *голова* ó *my head*(*your head, his head, her head*) etc."). Comparison of FDs demonstrate that Mongolian (like other Altaic languages) has moved further then European languages in the scale of grammaticalization of possessive pronouns. As an important result of this movement evolves a high degree of desemantization of of possessive pronouns: such pronouns play rather the role of definiteness marker, of topic marker or a substantivization marker, than the role of possessivity markers. But such process is own to all grammatical categories, so the fact of desemantization of the possessive markers prove its grammaticalized nature.

4.3. Frequency of grammatemes morphological tagsets in Mongolian

Tab. 3 represents frequencies of grammatemes morphological tagsets. Column A represents grammatemes; column B gives an explanation of the used notation.

As the study was done on the corpus with unresolved homonymy sometimes the status of morphological tagsets is ascribed not actually to grammatemes, rather to disjunctive bundles of (partially) homonymic grammatemes. Nevertheless the information on the frequency of such units in the corpus is not less important than the information on the frequency of real grammatemes. What was said above about the reason about effectiveness of use of the corpus with unresolved homonymy mutatis mutandis concerns also the grammar.

The fact that there are some disjunctive morphological tagsets in which the left member of a disjunction is identical with the right, can be explained by the fact that in ML there are many partially-homonymic pairs (a) in which the relation of subcategorial conversion (namely, one of members of pair belongs to thematic declination, and another — to athematic) takes place, and also (b) members of which differ in such a way that the final element of one of the members of such pair contains a steady «н», and the final element of the other member of the pair contains an unstable «н».

Table 3

A	B	C	D	E
NOM	nominative case	281693.26	885	1
0	the unique form of an uninflected word	93456.52	865	2
CVB.CNGR	gerund, congressive	50195.83	806	3
NOM~GEN		34055.37	857	4
PC.PROSP-NOM	participle, prospective	30514.26	794	5

A	B	C	D	E
CVB.MOD	gerund of manner	23820.50	733	6
NOM~CVB.MOD		23777.25	777	7
PC.PRF-NOM	participle, perfective, nom. case	23658.72	767	8
NOM~ABS-NOM		23602.48	753	9
GEN	genitive case	16387.03	779	10
NOM~VF.OPT.IMP		12322.50	675	11
DAT	dative case	11860.50	753	12
VF.IND.AOR	predicative role, indicative mood, aorist	11502.33	489	13
POSS.REFL	reflexive-possessive	10403.57	664	14
NOM~DAT		9257.23	639	15
REL-NOM	nominative case, attributive role	9122.27	662	16
NOM~REL-NOM		8674.98	677	17
NOM~COM-NOM		8229.42	646	18
NOM~VF.OPT.JUSS		8070.23	666	19
CVB.ANT	gerund, antecessive	8031.30	548	20
PC.PROSP-DAT	participle, prospective, dative case	7071.83	490	21
VF.IND.PRS1	predicative role, indicative mood, present № 1	6936.00	590	22
GEN/ACC	truncated form of genitive-accusative	6495.64	573	23
ACC	accusative case	6383.16	647	24
ABL	ablative case	4746.28	532	25
PC.US-NOM	participle, usual, nominative case	4708.21	477	26
0~VF.OPT.IMP		4624.29	526	27
A.COM-NOM~DAT		4610.45	552	28
NOM~ACC~VF.OPT.JUSS		4435.69	377	29
VF.OPT.IMP	predicative role, optative mood, imperative	4057.61	457	30
NOM~DAT~CVB.ANT		3978.88	329	31
NOM~CVB.CNGR		3920.05	401	32
NOM~CAR-NOM		3806.71	461	33
NOM~GEN~GEN		3646.66	478	34
NOM~PC.PRF-NOM		3352.50	534	35
VF.IND.PROF-EMPH	predicative role, indicative mood, profect, emphatic	3308.38	370	36
NOM~PC. PROSP-CAR-NOM		3278.10	390	37

A	B	C	D	E
NOM~A.COM-NOM~DAT		3273.77	509	38
COM-NOM	comitative case, nominative case	3156.98	415	39
NOM~CVB.ANT		3146.59	363	40
DAT~CVB.ANT		3114.58	390	41
NOM~CVB.TERM		3023.74	369	42
CVB.MOD~PC.PRF-NOM		2951.93	365	43
INSTR	instrumental case	2909.54	449	44

A lot of useful information can be drawn based on top-lists of grammatemes in ranging dictionaries. Such information gives impulse to some interesting observations of typological nature. Below we attempt to present some of such observations.

1. Gerundial forms of verbs (so-called converbs) occupy high positions in the FD of grammatemes: e. g., congressive converbs — 50195.83 ipm, modificative converbs — 23820.50 ipm, antecessive converbs — 8031.30 ipm.

In West European (Romanic and Germanic) languages (unlike Uralic and Altaic) converbs are used with lower frequency. Why such difference? The difference can be explained in the following way. Mongolian is characterized by the so-called Altaic type of coordination that means a rich system of gerundial taxis constructions in the domain of verb. Thus, it prefers the asyndetic coordination. Rare use of conjunctions is being compensated with a rich system of converbs.

2. Reflexive-possessive forms are used extraordinary frequently ("bare" reflexive-possessive forms — 10403.57 ipm; reflexive-possessive forms of the perfective participles — 2336.80 ipm; reflexive-possessive forms of the dative case of the prospective participles — 2001.12 ipm, etc.). The most familiar for us European languages do not have such form at all. The Russian does have the nearest translational equivalent of it - the reflexive-possessive pronoun *свой* (literally, 'his own', 'its own', 'her own', 'my own' etc., or 'of himself', 'of herself', 'of myself' etc.). But if we compare the relative frequency of *свой* with the relative frequency of the reflexive-possessive forms, we see that *свой* is used with a lower frequency. Its relative frequency is 3825.5 ipm.

Why such difference? The difference can be explained in the following way. The reflexive-possessive forms of Mongolian are used as one of the main devices of expressing co-reference. Languages with articles prefer other way of expressing co-reference (mainly, definite articles). Russian prefers the so-called "zero" forms of expressing co-reference. In fact they have rather prosodic (supra-segmental) than "zero" nature; but the prosodic devices are expressed in the written form of language very rarely.

References

1. *Baranova V. V, Saj S. S.* Ot sostavitelej (From the composers), in Issledovanija po grammatike kalmytskogo jazyka (Researches on the grammar of Kalmyk language). (= Acta linguistica petropolitana = Trudy instituta lingvisticheskih issledovanij (Works of Institute of linguistic researches), vol.V, pt. 2) SPb.: Nauka, 2009, p. 7–21.
2. *Kasevich V. B.* Vvedenie v jazykoznanie (Introduction to linguistics). M.-SPb.: Akademija, 2011, 230 pp.
3. *Shajkevich A. Ja.* Kolichestvennye metody v lingvistike (Quantitative approach in linguistics) In Lingvisticheskij èntsiklopedicheskij slovar' (Linguistic encyclopedic dictionary). M: Sovetskaja Èntsiklopedija (Soviet Encyclopedia), 1990, p. 231.
4. *Ljashevskaja O. N., Sharov S. A.,* Chastotnyj slovar' russkogo jazyka (na materialah Natsional'nogo korpusa russkogo jazyka). (Frequency dictionary of modern Russian (on materials of the National corpus of Russian)). M: Azbukovnik, 2009. (= http://dict.ruslang.ru/freq.php)
5. *Croft B.* Typology and universals. Cambridge, 2003.
6. *Lehmann Ch.* Directions for interlinear morphemic translations. In Folia linguistica, 1982, Vol. 16, p. 193–224.
7. *Sharoff, Serge,* Meaning as use: exploitation of aligned corpora for the contrastive study of lexical semantics. In Proc. of Language Resources and Evaluation Conference (LREC02). May, 2002, Las Palmas, Spain, 2002.

СЛОВА СО ЗНАЧЕНИЕМ «ЛЕКАРСТВЕННОЕ ВЕЩЕСТВО»: ВЧЕРА И СЕГОДНЯ

Крылова Т. В. (ta-kr@yandex.ru)

Институт русского языка РАН, Москва, Россия

В этой статье рассматриваются слова, обозначающие лечебное вещество, а именно — *лекарство, препарат, средство*. Наша задача состояла в следующем: 1) дать полное, непротиворечивое и логичное толкование трем этим словам; 2) сформулировать дифференциальные семантические признаки, противопоставляющие их друг другу; 3) проследить, как изменилось значение рассматриваемых слов, а также соотношение между ними, за последнее время, в связи с развитием медицины и фармацевтики. В ходе исследования было обнаружено, что слово *лекарство*, которое изначально являлось доминантой этого синонимического ряда, в современном словоупотреблении имеет более узкое значение по сравнению с двумя другими синонимами и, в отличие от двух других слов, используется в достаточно ограниченном круге контекстов. (Так, *лекарство* обозначает лечебное вещество, предназначенное исключительно для лечения болезни, что для двух других слов не обязательно). Сужение употребления слова *лекарство* может объясняться тем, что оно возникло давно, существенно раньше, чем соответствующие значения у слов *препарат* и *средство*, поэтому особенности его использования отражают традиционные наивно-языковые представления в сфере фармацевтики. В связи с этим *лекарство* не выдерживает конкуренции со словами *препарат* и *средство*, имеющими более богатое значение и отражающими более современные научные представления, и на глазах вытесняется ими из современного словоупотребления.[1]

Ключевые слова: семантика, слова со значением лекарственного вещества, русская наивно-языковая картина мира

1. Введение

В этой статье мы хотели бы рассмотреть слова семантического поля «лечебное вещество». Соответствующее значение присутствует у слов *лекарство, препарат, средство* — оно реализуется в лексемах *лекарство 1, препарат 2* (в сочетаниях типа лекарственный *препарат, препарат*

[1] Доклад публикуется в сокращенном варианте. Полная версия доступна на сайте конференции «Диалог».

от аллергии), средство 3 (в сочетаниях *средство от головной боли, лекарственные средства*)[2] .

Все перечисленные слова (в интересующих нас значениях) имеют следующий общий элемент в толковании: «вещество, предназначенное для улучшения состояния человека или животного».

Толкования рассматриваемых лексем в словарях следует признать неудовлетворительными. В частности, в Малом академическом словаре *лекарство* толкуется как «Вещество, применяемое для лечения или предупреждения болезни», а *препарат* — как «Продукт лабораторного или фабричного изготовления лечебного или другого назначения», при этом неясно, чем вещество для лечения болезни отличается от продукта лечебного назначения . В связи с тем, что различие между *лекарством* и *медицинским препаратом* остается неясным, вызывает недоумение и толкование слова *средства* в интересующем нас значении, которое звучит следующим образом: «Лекарства, медицинские препараты или предметы, необходимые при лечении». Вся информация о различиях между рассматриваемыми лексемами, которую можно почерпнуть из этих толкований, сводится к тому, что *препарат,* в отличие от *лекарства* и *средства*, не обязательно предназначен для лечения, а *средство,* в отличие от *препарата* и *лекарства,* может обозначать предметы. Однако интуитивно мы ощущаем, что различия между этими словами лежат значительно глубже и не сводятся к описанным особенностям.

Поэтому первая из задач, которая ставится в данной статье, состоит в том, чтобы дать полное, непротиворечивое и логичное толкование трем интересующим нас словам, а также сформулировать дифференциальные семантические признаки, противопоставляющие их друг другу.

Далее, мы хотели бы проследить, как изменилось значение рассматриваемых слов, а также соотношение между ними, за последнее время, в связи с развитием медицины и фармацевтики.

Кроме того, мы попытаемся реконструировать тот фрагмент языковой картины мира[3] , который касается представлений о лечебном веществе, и проследить, как изменились языковые представления в этой сфере за последнее время.

---

[2] В дальнейшем, говоря о словах *лекарство, препарат, средство,* мы будем иметь в виду только те лексемы, которые имеют интересующее нас значение — а именно, значение лечебного вещества. У слова *средства* наличие отдельной лексемы с таким о значением не является бесспорным, однако словари выделяют это значение как самостоятельное. Так, Малый академический словарь (МАС) внутри значения «предметы, приспособления или совокупность их, необходимые для осуществления чего-л» выделяет подзначение «лекарства, медицинские препараты или предметы, необходимые при лечении» *(болеутоляющие средства, дезинфицирующие средства, перевязочные средства*). Русский толковый словарь под редакцией В. Лопатина идет по тому же пути: наряду со значением «орудие для осуществления какой-л. деятельности» он выделяет значение «лекарство, предмет, необходимый при лечении» (*средство от простуды, средство против гриппа, перевязочные средства).*

[3] О наивной картине мира и наивно-языковых представлениях см. в [Апресян 1995:56], [Урысон 2003:11–12].

Наконец, мы хотели бы в осторожной форме высказать некоторые прогнозы относительно того, какое из рассматриваемых слов получит преимущество в дальнейшем, исходя из его семантических и стилистических особенностей.

Описание интересующей нас группы слов будет построено в виде лексикографических «портретов», причем каждое новое слово мы будем рассмаривать в его противопоставленности двум другим словам.

2. Лекарство

Рассмотрим, в чем специфика семантики слова *лекарство* и чем оно отличается от *средства* и *препарата,* вместе взятых.

Начнем с того, что *лекарство*, как наиболее традиционное обозначение лечебного средства, является доминантой данного синонимического ряда. Это проявляется в том, что *лекарство* может обозначать лечебное вещество вообще, без конкретизации его характера и предназначения. Таковы употребления типа: *Он принял лекарство; Врач выписал лекарство; Сходи купи лекарство; Его увели в дом, уложили, дали лекарство* (Б. Екимов); *Она приходит домой с болью в сердце, боль усиливается, она вызывает "скорую помощь", пьёт лекарство, но тщетно, — и умирает* (А. Слаповский).

Для *препарата* и *средства* вышеназванный тип употреблений невозможен или нетипичен. Ср. странное *Он принял препарат*; *Врач выписал препарат; купить <дать> препарат* (все эти сочетания становятся нормальными, если добавить оределение: *препарат, выписанный кардиологом* или *препарат для снижения давления*). (Что касается *средства,* то у этой вокабулы интересующее нас значение вообще реализуется главным образом в сочетании с определением, реализующим валентность предназначения — *средство от насморка, противогрибковые средства).*

Итак, *лекарство* имеет в целом более широкое значение, по сравнению с *препаратом* и *средством.* С этим хорошо согласуется его стилистическая нейтральность, которая проявляется, в частности, в том, что *лекарство,* единственное из рассматриваемых слов, может использоваться в разговорной речи: ср. вполне стандартное *Куда ты засунула лекарство от кашля?; Он накупил кучу дорогущих лекарств от гастрита*. Два других слова являются необиходными (при этом *препарат* тяготеет к научному стилю, а *средство* — к официальному стилю речи), в связи с чем их использование в вышеприведенных контекстах выглядело бы неестестенно.

При этом парадокс в том, что в некоторых аспектах *лекарство* обнаруживает большую узость значения, чем два других синонима. Это во многом объясняется тем, что у *лекарства* интересующее нас значение (которое является для него основным) возникло давно, существенно раньше, чем соответствующие значения у слов *препарат* и *средство,* поэтому особенности его использования отражают традиционные наивно-языковые представления в сфере фармацевтики.

Основное различие между *лекарством* и двумя другими синонимами связано с характером функции, или назначения. *Лекарство* предназначено для

лечения болезни А1 или для лечения органа А1. Валентность А1 заполняется предложно-именными группами *от* + РОД и *против* +РОД (*лекарство от простуды <от кашля, от головной боли>, лекарство против аллергии, лекарство от головы <от горла, от сердца, от печени>)* или прилагательными (*желудочное <сердечное> лекарство)*.

Лекарство не используется, если в роли А1 выступает название какого-л. отклонения от нормы, которое не является названием болезни — ср. странное *лекарство от учащенного сердцебиения* при нормальном *лекарство от тахикардии;* странно также *лекарство от высокой температуры* при нормальном *лекарство от лихорадки <гриппа>*.

При этом употребление слова *лекарство* дает возможность проанализировать, какие именно виды патологий связываются в языковом сознании с представлением о болезни. Приходится признать, что *кашель* — самостоятельная болезнь — на это указывают вполне стандартные сочетания *лекарство от кашля,* которые существуют наряду с сочетаниями типа *лекарство от бронхита*. Судя по допустимости соответствующих сочетаний, согласно языковым представлениям, болезнью также является *головная боль* и *бессоница:* ср. *Все хорошо. Мне нужно лекарство от головной боли* (И. Муравьева); *Нормальному человеку нельзя просто купить лекарство от бессонницы* («Частный корреспондент», 2010).

Кроме того, судя по употреблению слова *лекарство,* такие виды патологий, как *алкоголизм, наркотическая зависимость,* с наивно-языковой точки зрения, болезнями не являются — на это указывает нестандатрность сочетаний типа *лекарство от алкоголизма <от наркотической зависимости>* (такие сочетания возможны, если *лекарство* употреблено в переносном значении: ср. *Единственное настоящее лекарство от алкоголизма — сделать жизнь такой, чтобы от нее не хотелось опохмеляться* («Новый мир», 2000). Между тем, *средство* и *препарат* в этом случае вполне уместны: ср. *Не поддавайтесь на увещевания разного рода знахарей, которые рекомендуют к применению, например, чемерицу (Veratrum) или копытень (Asarum) как средство от алкоголизма, — эти растения очень ядовиты* (Сад своими руками», 2002.12.15); *Где купить эффективный препарат от пьянства «АЛКОПАН»?* («Аргументы и факты», 2001.02.14).

Теперь обратимся к словам *препарат* и *средство.* Их назначение мыслится несколько иначе, чем у слова *лекарства*. Во-первых, оно часто связывается не с лечением болезни (хотя и это возможно: ср. *противотуберкулезный препарат, противотуберкулезное средство, препарат <средство> от аллергии)*, а с конкретным воздействием на организм или орган *(жаропонижающий <дезинфицирующий, стимулирующий, имунный> препарат; противогрибковое <обезболивающее, успокоительное> средство).* Для *лекарства* большинство сочетаний такого типа затруднены; ср. нестандартное *жаропонижающее <дезинфицирующее, стимулирующее, обезболивающее, успокоительное, имунное> лекарство.*

Указание на характер воздействия в случае *препарата* и *средства* может вводиться не только с помощью прилагательного, но и с помощью

предложно-именной группы *для* + РОД, где в форме РОД используется существительное с названием типа воздействия. Приведем следующие типичные сочетания: *препарат <средство> для улучшения пищеварения <для укрепления иммунитета, для лечения кашля, для профилактики гриппа, для снижения веса>*.

Важная особенность *препарата* и *средства,* отличающая их от *лекарства,* состоит в том, что они могут использоваться, когда воздействие не носит лечебного характера и не связано напрямую с медициной. Оно может быть направлено на достижение практических целей — гигиенических, косметических, спортивных, а также таких, как контрацепция и пр. Ср. *очищающие средства <препараты>, косметические препараты <средства>; средства контрацепции, препараты <средства> для снижения веса; допинг-препараты; препараты для роста мышечной массы.* При этом интересно, что одно и то же вещество может быть названо как *лекарством*, так и *препаратом*, в зависимости от того, что находится в фокусе внимания — лечение болезни или другие практические цели; ср. *Астму лечат сильными лекарствами — в частности, стероидами; В спорте применяют сильные препараты* (но не *лекарства*)*— в частности, стероиды.*

Таким образом, *средство* и *препарат,* как слова, развившие значение лечебного вещества относительно недавно, в эпоху развития медицины и ее проникновения в другие сферы, демонстрируют размывание представления о назначении лечебного вещества исключительно для борьбы с болезнями. Между тем, именно это представление, сохранившееся в семантике слова *лекарство,* для данного семантического поля является первичным и прототипическим.

Наконец, последнее различие между *лекарством*, с одной стороны, и *препаратом* и *средством,* с другой, связано с типом вещества и со способом его введения в организм человека.

Лекарство чаще всего принимают внутрь (ср. стандартное *принять <выпить> лекарство*) или вводят в виде инъекции (ср. нормальное, хотя и чуть менее распространенное *ввести лекарство, колоть лекарство*). Значительно реже *лекарства* используют наружно — не вполне стандартно звучит *помазать <обработать> лекарством, нанести лекарство,* чуть лучше *положить лекарство, промыть лекарством*.

Описываемое свойство *лекарства* проявляется даже не столько в ограничениях на сочетаемость с глаголами, сколько в нетипичности употребления *лекарства* применительно к различным лечебным веществам, применяемым наружно. *Лекарством* мы обычно называем таблетки или микстуры; растворы для наружного применения и мази в современном языке обычно не называют *лекарствами* — их естественне назвать *средствами* или *препаратами*. Ср. странное *Мне посоветовали купить одно лекарство — мазь Левомеколь <Фастум-гель>; Перекись — лекарство, абсолютно необходимое в каждом доме.* Первая фраза становится абсолютно нормальной, если заменить *лекарство* на *средство* или *препарат*, во второй естественнее выглядит *средство* (о причинах этого см. ниже).

При этом лекарства обычно предназначены не просто для введения внутрь организма, а для попадания в желудок — так, ингаляторы для горла,

предназначенные для нанесения на слизистую оболочку горла, *лекарством* вряд ли назовут.

Итак, уточненное толкование *лекарства* звучит следующим образом. ЛЕКАРСТВО — **«специально созданное вещество, которое предназначено для лечения болезни А1 или органа А1 и которое обычно используется для внутреннего применения».**

3. Препарат

Рассмотрим теперь слово *препарат*.

За последние десятилетия значение этого слова претерпело существенные изменения. Изначально оно означало «специальным образом обработанная часть А1 живого организма, предназначенная для изучения его устройства или для проведения научных экспериментов»: *анатомический препарат; учебный препарат; Здесь стояли [...] два бронзовых бюста неизвестно кого, несколько больших стеклянных банок с препаратами мозга — застывшими завитками цвета детского мыла* (Л. Улицкая); *Затем несколько раз побывал на выставках при женских гигиенических курсах, где были расставлены и развешаны муляжи человеческих органов, а также заспиртованные препараты* (П. Нилин).

Еще в Малом академическом словаре это значение подается в качестве исходного, тогда как значение вещества, основное для современного словоупотребления, расценивается как производное и получает помету *хим., фарм.* («Продукт лабораторного или фабричного изготовления лечебного или другого назначения»).

За последнее время значение вещества вышло у *препарата* на первый план, что сопровождалось ослаблением стилистических ограничений на его употребление (сейчас *препарат* ощущается не как специальное, а как необиходное слово). Кроме того, *препарат* как вещество стал употребляться в двух разных значениях.

1) «Специально приготовленное вещество, используемое в сфере А1» (*препарат, используемый для хранения крови <для борьбы с вредителями>; химический препарат*). Ср. также *В настоящее время разрабатываются препараты для многофункциональной отделки целлюлозосодержащих тканей* («Текстиль», 2003).

2) «Специально приготовленное вещество, используемое для улучшения состояния живого существа или его органов». (Это толкование является приблизительным и впоследствии будет уточнено). Данное значение реализуется в сочетаниях типа *лекарственный <медицинский, фармацевтический> препарат; сильнодействующий <действенный> препарат; ввести препарат; дозировка препарата*. Многие из этих сочетаний возможны только для «медицинского» *препарата:* ср. странное *сильнодействующий <действенный> химический препарат <препарат для борьбы с вредителями>.*

В последнее время “медицинское» значение стало для *препарата* основным. В этом значении *препарат* постепенно вытесняет *лекарство.* Именно его мы будем рассматривать в дальнейшем.

Препарат отличается от других рассматриваемых слов тем, что обнаруживает связь с научной сферой — по-видимому, это свойство унаследовано интересующим нас значением *препарата* от исходного «анатомического» значения.

Поскольку *препарат* тяготеет к научной сфере, он акцентирует внимание на происхождении вещества, используемого для лечения. Вследствие этого для *препарата* характерно указание на химический состав вещества; в отличие от *лекарства* и *средства* он имеет валентность химического элемента. Эта валентность заполняется формой РОД, сочетанием предлога *на основе* + РОД или прилагательным: ср. *препараты железа, препараты на основе свинца, свинцовые препараты; Он [...] ввел в употребление препараты ртути, благодаря чему эпидемию удалось остановить* («Наука и жизнь», 2007); *Полный покой, абсолютное молчание в течение нескольких дней, бромистые препараты, витамины* (В. Шишкин). Между тем, у *лекарства* и *средства* валентность химического элемента отсутствует — ср. невозможное *лекарство <средство> железа, свинцовые лекарства <средства>*.

Наконец, специфика *препарата* состоит в том, что он является достаточно сложным, тогда как *средство* может быть простым по составу. Вещества типа *перекиси, зеленки, спиртового раствора календулы* едва ли могут быть названы *препаратами,* тогда как назвать их *обеззараживающими средствами* вполне нормально.

Итак, окончательное толкование *препарата* звучит следующим образом. ПРЕПАРАТ — **«специально приготовленное вещество, имеющая сложный состав, основой которого является химический элемент А1, предназначенное для лечения болезни или органа А2 или оказания воздействия А2 на организм или орган».**

4. Средство

Перейдем к слову *средство. Средство,* в отличие от лекарства и препарата, не обязательно специально создано человеком; это может быть вещество, существующее в природе, а также часть растения и пр. (Так, *мед, малина, морская соль* могут быть названы лекарственными средствами, но не *лекарствами* и не *препаратами*). Ср. *Индейцы Центральной и Южной Америки традиционно использовали плоды кокосовой пальмы как противовоспалительное средство* («Наука и жизнь»).

Кроме того, *средство,* в отличие от *лекарства* и *препарата,* может обозначать не только продукт традиционной медицины, созданный в условиях фабричного или лабораторного производства, но и продукт, созданный в домашних условиях. По этой причине нормально звучит *народное средство,* но странно *народное лекарство, народный препарат*.

Далее, *средство,* в отличие от *лекарства* и *препарата,* не обязательно представляет собой вещество. Это может быть также предмет, используемый для лечения (*перевязочные средства).* Показательно также различие между сочетаниями *противозачаточный препарат* и *противозачаточное средство:* если первое обозначает особое вещество, которое обычно выпускается в виде

таблеток, то второе может обозначать также особые приспособления, в частности, спираль и пр.

Еще одно отличие *средства* от *препарата* и *лекарства,* вместе взятых, касается характера предназначения. *Средство,* в отличие от двух других слов, может быть предназначено для устранения различных нежелательных явлений частного характера, являющихся проявлением плохого состояния того или иного органа или организма в целом — ср. *средство от боли <выпадения волос, появления морщин, перепадов давления, быстрой утомляемости, отеков ног>.* Именно с этой особенностью *средства* связана его невероятная активность в рекламе. Ср. *«Стрепфен» — средство от боли в горле с энергетическим действием* (Реклама на канале ОРТ (2006)); *«Пенталгин-Н» / средство от боли / содержит пять активных компонентов* [Реклама на ТВ (2006).

Наконец, *средство,* в отличие от *лекарства* и *препарата,* свободно обозначает любой тип вещества и допускает любой тип введения в организм — его можно принимать внутрь *(принять средство от мигрени <средство для лечения кашля, средство, разжижающее мокроту>),* вводить в виде инъекций (*ввести средство от аллергии <успокоительное средство>)* и, главное, наносить наружно. В связи с этим *средство* охватывает собой все лечебные вещества, употребляемые наружно — мази, растворы для наружного применения и кремы; отсюда его распространенность в косметологии и рекламе. Ср. *средство от морщин <от жирной кожи, от угревой сыпи, от пигментных пятен, от мозолей, от псориаза, от облысения>.*

Возможность свободно употребляться применительно к веществам, используемым наружно, отличает *средство* не только от *лекарства,* но и от *препарата.* Для последнего такое употребление, хотя и допустимо, но не очень характерно. Описываемое различие между *препаратом* и *средством* хорошо иллюстрирует минимальная пара *препарат для омоложения кожи* и *средство для омоложения кожи:* если первое сочетание, скорее всего, обозначает вещество, вводимое в виде инъекций, то второе — скорее всего, крем.

Окончательное толкование слова *средство* в интересующем нас значении выглядит так: СРЕДСТВО — **«вещество или предмет, используемые для лечения болезни или органа А1, для оказания воздействия А1 на организм или орган или для устранения нежелательного явления А1, являющегося проявлением плохого состояния организма или органа»**. Таким образом, пожалуй, именно *средство* в современном языке является наиболее универсальным словом, обозначающим лечебное вещество — не случайно в аптеках именно это слово используется для классификации лекарств (всем знакомы таблички *«средства от кашля», «средства от насморка», “имунные средства», «жаропонижающие средства»).*

5. Выводы

Перечислим дифференциальные признаки, различающие слова в рамках рассматриваемого ряда.

	Лекарство	**Препарат**	**Средство**
1.Характер предназначения	Лечение болезни	Лечение болезни; конкретное воздействие на какой-л. орган	Лечение болезни; конкретное воздействие на какой-л. орган; устранение какого-л. нежелательного явления частного характера
2.Способ введения вещества в организм	Внутреннее применение	Чаще внутреннее применение, чем наружное	Внутреннее и наружное применение
3. Наличие указания на химический состав вещества	–	+	–
4. Степень сложности вещества	Не охарактеризовано	Сложное вещество	Не охарактеризовано
5.Происхождение вещества	Искусственное происхождение	Искусственное происхождение	Искусственное или природное происхождение
6. Возможность обозначать не только вещество, но и предметы, используемые при лечении	–	–	+
7. Сфера употреблени: возможность использования не только в сфере традиционной медицины, но и в сфере народной медицины	–	–	+

Итак, слово *лекарство,* некогда являвшееся основным в семантическом поле «лечебное вещество», на наших глазах теряет свое преимущество, не выдерживая конкуренции со словами *препарат* и *средство,* имеющими более богатое значение и отражающими более современные научные представления. Два этих слова обнаруживают приблизительно одинаковую активность, причем каждое из них заняло отдельную нишу: *препарат* активен в сфере науки, а *средство* — в сфере потребления (и особенно в рекламе). При этом можно предположить, что рано или поздно *средство* займет центральное место в данном семантическом поле, поскольку является более универсальным как

по своей семантике (что наглядно показывает приведенная выше таблица), так и по своей стилистической принадлежности.

References

1. *Apresian Iu. D. (1995) Lexical semantics [Leksicheskaia semantika] Moscow, Languages of Slavonic culture.*
2. *Uryson Ie. V. (2003) Problems of research of nanve world model [Problemy issledovaniia iazykovoi kartiny mira] Moscow, Languages of Slavonic culture.*

СЕМАНТИЧЕСКИЕ МОДЕЛИ И СЕМАНТИЧЕСКИЕ ФУНКЦИИ ОТПРИЧАСТНЫХ ПРИЛАГАТЕЛЬНЫХ

Кустова Г. И. (galinak03@gmail.com)

Московский педагогический государственный университет, Москва, Россия

Рассматриваются типы связей адъективированных причастий (отпричастных прилагательных) с глагольным событием (ср. *играющий тренер*, *приглашенный профессор*, *воспитанный юноша*) и типы импликаций, на основе которых развиваются качественные значения отпричастных прилагательных (*повышенная температура*, *распространенное мнение*, *продуманное решение*).

Ключевые слова: семантика, прилагательное, причастие, адъективация, импликация, качественное значение

В работе будут рассмотрены семантические группы адъективированных причастий.

Адъективированное причастие (ср. *играющий тренер*), в отличие от настоящего причастия, которое может обозначать событие вместо глагола (ср. *Мы увидели тренера, игравшего в футбол / играющего в футбол вместе с командой*), утрачивает глагольные признаки вида и времени. Именно эти процессы обычно находятся в центре внимания исследователей [Лопатин 1966; Калакуцкая 1971; PP 1973: 177–185]. Другой важный процесс при переходе в «адъективное» состояние — утрата глагольных актантов.

Мы будем делать акцент не на том, что адъективированные причастия — это бывшие причастия, а на том, что это (нынешние) прилагательные. Поэтому в дальнейшем мы будем называть их отпричастными прилагательными и сравнивать с другими отглагольными прилагательными. Событие, обозначаемое производящим глаголом, мы будем называть глагольным событием, или просто событием. А определяемое существительное — просто существительным, или предметом.

Хотя отпричастное прилагательное перестает быть глагольной формой, оно сохраняет какую-то связь с глагольным событием (за исключением случаев полной утраты такой связи типа *блестящий музыкант*). Ниже мы рассмотрим, какую роль играет глагольное событие в адъективном значении причастия.

Самыми многочисленными из всех групп отпричастных прилагательных являются две — от активных причастий настоящего времени НСВ (*кормящая мать*, *лечащий врач*, *руководящий работник*, *играющий тренер*, *моющее средство*, *увлажняющий крем*, *пишущая машинка*, *записывающее /*

подслушивающее устройство, *бегущая дорожка*, *пьянящий запах*, *давящая повязка* и под.) и от пассивных причастий прошедшего времени СВ (*смешанный лес*, *сосредоточенный огонь*, *организованный туризм*, *запрещенный препарат*, *защищенная линия связи*, *обманутые вкладчики* и под.). Они не просто самые многочисленные, но и постоянно пополняемые. Это говорит о том, что эти две группы — точки роста, важные для говорящих, это поставщики недостающих, нужных для коммуникации средств обозначения признаков. Первую группу мы условно назовем активными отпричастными прилагательными, хотя далеко не всегда эти отпричастные прилагательные имеют агентивный компонент, ср. *пьянящий запах*; вторую условно назовем перфектными отпричастными прилагательными. Ниже будут рассматриваться группы перфектных отпричастных прилагательных, активные послужат для них лишь фоном.

Большинство активных отпричастных прилагательных — относительные прилагательные. Относительные прилагательные содержат отсылку к чему-л.: отсубстантивные — к предмету (ср. *книжный* — отсылка к книгам: *книжный магазин*, *книжный шкаф*), отглагольные — к действию (*читальный зал* — ‘где читают’). В обоих случаях есть предикат, связывающий участников отношения. Но в первом случае он «вычисляется» говорящими (*книжный магазин* — ‘где продают книги’; *книжный шкаф* — ‘где хранят книги’), во втором предикат (событие) указывается эксплицитно, а вычисляется второй участник отношения (*увлажняющий крем* — ‘увлажняет кожу’; то же относится к относительным перфектным отпричастным прилагательным: *приглашенный профессор* — ‘пригласил тот, кто имеет полномочия’).

Для большинства активных отпричастных прилагательных и многих других отглагольных прилагательных глагольное событие — это действие, которое многократно повторяется либо может быть (обычно тоже многократно) совершено: характеристика человека содержит действие, которое он регулярно производит: *играющий тренер*, *кормящая мать*, *пьющий сосед* (это похоже на nomina agentis — свойство, которое состоит в повторяющемся действии: *учитель*, *дирижер*, *рыбак*, *коллекционер*); характеристика предмета обычно содержит его функцию: *режущий инструмент*, *освежающие пастилки*. Даже экспериенциальные отпричастные прилагательные не отличаются от «агентивных» с точки зрения узуальности (или потенциальности): *пьянящий запах*, *бодрящий аромат кофе* — пьянит / бодрит любого наблюдателя, который чувствует этот запах. Таким образом, у относительных активных отпричастных прилагательных содержание сводится, в основном, к глагольному событию.

ЗАМЕЧАНИЕ.

Качественных активных отпричастных прилагательных довольно мало, и мы вернемся к ним, когда будем обсуждать качественные перфектные отпричастные прилагательные ОП.

Если большинство активных отпричастных прилагательных соотносятся с реальными активными причастиями, то перфектные отпричастные прилагательные далеко не всегда соотносятся с реальными пассивными причастиями от переходных глаголов. Например, *приглашенного профессора* действительно

пригласили, но *учащенный пульс* не участили — он сам участился. Вообще, среди перфектных отпричастных прилагательных достаточно много форм, соотносимых с возвратными глаголами (типа *учащенный*, *искривленный*, *вытянутый*), встречаются также более сложные случаи псевдопричастных образований: *заспанные глаза*, *восторженный поклонник*, *воздержанный образ жизни*, *предубежденный взгляд*, *начитанный школьник*, *ухоженные руки* и т. п. Мы условно будем считать все такие формы отпричастными, т. к. их морфемная структура формально соответствует структуре страдательного причастия СВ. Для нас существенно также, что они соотносятся не просто с глагольной основой, но с реальным или воображаемым глагольным событием. При этом у многих из них есть синонимичные или просто близкие по значению «обычные» прилагательные, которые, в принципе, могут обозначить то же состояние или признак (*брошенные дома — пустые*; *запрещенные препараты — вредные*, *опасные*; *заспанное лицо — сонное*; *упрощенная процедура — простая*). Тем не менее у говорящих есть потребность иметь еще и аналогичное отпричастное прилагательное. Это значит, что отпричастные прилагательные выражают какой-то важный признак, который может выразить именно эта семантическая структура.

Что общего у всех перфектных отпричастных прилагательных — как относительных, так и качественных? Несколько огрубляя, можно сказать, что это ретроспективная отсылка.

Активные отпричастные прилагательные содержат проспективную отсылку — к событию, которое может или будет происходить, как правило многократно. Конечно, такие признаки, как *пьющий сосед*, *играющий тренер* и т. п., обобщают действие, которое происходило раньше (на достаточно протяженном интервале). Существенно, однако, то, что такое действие можно ожидать от данного лица (или предмета) и в будущем (иначе оно не рассматривалось бы как свойство), и характеристикой лица (предмета) является не только обычность, но и возможность такого действия.

Перфектные отпричастные прилагательные содержат отсылку к прошлому событию, чаще однократному (*брошенные дома* — их бросили жители), но, возможно, и длительному или повторявшемуся (*изношенное пальто* — износилось в результате длительной носки; *обветренное лицо* — обветрилось в результате, возможно, многократного воздействия ветра). Причем говорящих интересует не столько само это событие, сколько его последствия для объекта (в этом состоит перфектность, результативность — не только в сфере отпричастных прилагательных, но и в сфере собственно глагольных форм перфекта и результатива, см. [Типология 1983]).

Поскольку перфект и результатив — это состояния, наступившие в результате прошлого действия, то для перфектных отпричастных прилагательных важно включить в семантику этого состояния некую предысторию. Сравним выражения *брошенные дома* и *незаселенные дома*. Денотативно это одно и то же — ‘пустые, без жильцов’. Но признак *пустые* содержательно бедный, он не включает никакой предыстории, никакого объяснения этого

денотативного состояния. А *брошенные* и *незаселенные* содержат предысторию, причем разную.

Отсылка к глагольному событию часто важна не сама по себе, а из-за тех возможностей, которые она предоставляет говорящим. Рассмотрим эти возможности и группы отпричастных прилагательных, которые формируются в результате их реализации (ниже приводятся лишь наиболее репрезентативные группы отпричастных прилагательных, список их, разумеется, не исчерпывающий).

МОТИВИРУЮЩЕЕ СОБЫТИЕ (*приглашенный профессор*).

К этой группе относятся, в частности, такие случаи: *перемещенные лица*, *приглашенный профессор*, *прикрепленные сотрудники*, *разрешенный митинг*, *согласованное шествие*, *привлеченные средства*, *упущенная выгода*, *отложенный спрос / платеж*; *объединенный народный фронт*, *запрещенные препараты*.

Событие, к которому отсылает отпричастное прилагательное данной группы, никак не влияет на «видимое» состояние предмета, его нельзя реконструировать по внешнему виду предмета, если заранее о нем не знать. Если бы это событие V не произошло, X нельзя было бы назвать *V-енным*, с другой стороны, по тому, каков X, обычно нельзя установить, что событие V имело место.

Так, *запрещенные препараты* могут быть вредными, опасными для человека, но *запрещенными* они являются только потому, что их запретили. В противном случае при тех же свойствах они не были бы запрещенными. Запрет — это событие в информационном пространстве. О нем нельзя узнать по цвету или химическому составу препарата. Его также нельзя вывести из ненаблюдаемых, но имеющих реальные, наблюдаемые проявления свойств *вредный* или *опасный*. Многие сердечные средства являются вредными для желудка и печени, тем не менее это *разрешенные препараты*. Событие V, которое объясняет, почему предмет так называется, обозначим как мотивирующее событие.

Зачем нужна отсылка к мотивирующему событию, если она ничего не дает для понимания настоящего, а связь с этим событием внешне ни в чем не выражается? Отпричастные прилагательные данной группы являются относительными прилагательными. Большинство относительных прилагательных обозначает признак, который выделяет подкласс в заданном существительным классе, ср. подклассы часов: *наручные часы*, *настольные часы*, *настенные часы* и т. п., подклассы супов: *грибной суп*, *мясной суп*, *овощной суп* и т. п. (см. [Павлов 1960], [Панов 1999]). Наручные и настольные часы принципиально, сущностно не отличаются — те и другие показывают время, а отличаются лишь по внешним, второстепенным признакам. Тем не менее это разные виды часов. То же самое — с мотивирующим событием. *Приглашенный профессор*, например, — это подкласс профессоров. «Внешне» он не отличается от других профессоров, т. к. точно так же читает лекции студентам. От другого подкласса — штатных профессоров — приглашенные отличаются, строго говоря, не тем, что их пригласили, а тем, что они работают временно, а не постоянно, как штатные. Таким образом, мотивирующее событие задает признак, может быть, не столь уж яркий и существенный, но достаточный, чтобы выделить подкласс из общего класса.

«НАСТОЯЩИЙ» ПЕРФЕКТ (*окрашенные волосы*).

Есть группа отпричастных прилагательных, которая отсылает к значимой предыстории состояния — т. е. к ситуации, которая на протяжении какого-то времени воздействовала на объект и меняла его реальное, часто физическое состояние. Это и есть традиционно понимаемое перфектное значение — состояние (изменение) в результате предшествующего события, воздействия (точнее, группа значений — перфект, результатив и под., см. [Типология 1983]), ср. *окрашенные волосы*, *воспаленные глаза*, *заспанные глаза*, *обветренное лицо*, *обмороженные руки*, *изношенные ботинки*. Такие отпричастные прилагательные часто имеют аналоги среди бесприставочных отглагольных прилагательных типа *крашеный*, *стриженый*, *мороженый*, *сушеный*, *ношеный*, *рваный* и под. (мы, в соответствии с традицией, считаем единицы типа *вареный*, *мороженый*, *рваный* не причастиями, а прилагательными, т. к. их форма (соотнесенность с основой НСВ) не соответствует значению, содержащему отсылку к завершенному действию СВ ‘сваренный’, ‘замороженный’, ‘порванный’). Однако перфектные отпричастные прилагательные не являются лишними, т. к. концептуализуют другой аспект того же или близкого денотативного содержания. Так, *крашеные волосы* означают просто, что их цвет не «свой», не натуральный. А признак *окрашенные* подчеркивает, что в результате окрашивания их свойства могли измениться (и для восстановления их исходных природных свойств предназначены, например, специальные шампуни «для окрашенных волос»).

Показательно, что часто и отпричастное прилагательное и глагол в перфектном значении обозначают более или менее то же самое состояние, ср. *У него глаза воспалились — У него воспаленные глаза*; *У него лицо обветрилось — У него обветренное лицо*.

Наибольший интерес для нас представляют случаи, когда отпричастное прилагательное подразумевает не реальную предысторию, а создает интерпретацию наличного признака, моделируя с помощью причастной формы событие, которого в реальности могло и не быть. Наличное состояние выглядит так, как будто оно является следствием предшествующего события. Но для говорящего важно не столько событие, сколько следствие. Рассмотрим три таких случая.

КОМПАРАТИВНАЯ МОДЕЛЬ (*расширенная программа*).

Это большой класс отпричастных прилагательных, состоящий из двух основных подклассов. Первый подкласс включает отпричастные прилагательные от глаголов, в свою очередь образованных от прилагательных (иногда — от компаративов типа *лучше*, *меньше*): *завышенные требования*, *замедленное развитие*, *заниженные показатели*, *затемненные стекла*, *затрудненное дыхание*, *облагороженный вид*, *облегченная конструкция*, *обновленная версия*, *осложненная травма*, *осовремененная постановка*, *охлажденный кофе*, *повышенные запросы*, *пониженная температура*, *расширенная программа*, *сниженные цены*, *сокращенный вариант*, *улучшенная планировка*, *уменьшенный вид*, *упрощенная процедура регистрации*, *усложненные правила регистрации*, *убыстренный темп*, *увеличенный рабочий день*, *углубленное изучение*, *укороченный рукав*,

уплотненный картон, *усиленное питание*, *ускоренная процедура регистрации*, *утепленное пальто*, *учащенное сердцебиение* и под. К ним примыкают некоторые неотадъективные слова, семантически построенные по той же компаративной модели: *модернизированный проект*, *ограниченный доступ, сдержанный оптимизм, сжатые сроки*, *сжиженный газ*, *щадящий режим*. Эти отпричастные прилагательные обозначают изменение значения некоторого параметра.

ЗАМЕЧАНИЕ. Такие прилагательные, как *обостренный слух* (vs. *заостренная форма*), *утонченный вкус*, *возвышенные мысли* имеют другую семантическую модель и к этой группе не относятся.

Второй подкласс обозначает «измененную» форму: *вытянутый подбородок*, *закругленный кончик*, *заостренный кончик*, *изогнутый меч*, *искаженное изображение*, *искривленный палец*, *ссутуленная спина*, *удлиненное лицо*, *уплощенная стопа*, *утолщенный стебель* (многие такие отпричастные прилагательные, как и в первом подклассе, генетически связаны с прилагательными, ср. *круглый*, *кривой*, *длинный* и т. п., но здесь другой состав «предметов» — это физические объекты, часто — «естественного происхождения»: части тела, части растений).

Описываемое этими отпричастными прилагательными состояние может быть, а может и не быть результатом некоторого события: *повышенная температура* действительно повысилась (была нормальная); *упрощенная процедура* оформления кредита не обязательно означает, что она была сложная, а потом ее упростили. Это не исключено, но возможно и другое: она проще, чем в других банках и чем могла бы быть в этом банке. Наконец, *искривленный ствол* никто не искривлял и тем более *вытянутый подбородок* никто не вытягивал, он просто отклоняется от обычного, стандартного, нормального (т. е. прямого, округлого).

Глагольное событие (реальное или воображаемое), к которому отсылают рассматриваемые отпричастные прилагательные, создает условную точку отсчета, с которой сравнивается наличное, наблюдаемое состояние и по отношению к которой это состояние интерпретируется как немного сдвинутое, смещенное. *Удлиненный* — это такой, как если бы его удлинили или он удлинился. Такие глаголы (Е. В. Падучева называет их градативами, см. [Падучева 2004]) обозначают некоторый сдвиг по оси признака или параметра (длины, температуры, кривизны, скорости). Т.е. отсылка к такому событию создает компаративную структуру, например:

упрощенная процедура — относительно простая, несложная, проще, чем могла бы быть; такой признак задает отношение к другой точке шкалы данного параметра. Признак, выражаемый прилагательным, — *простая процедура*, — подразумевает, скорее, противоположный полюс (*сложная процедура*);

удлиненный стебель — длиннее, чем мог бы быть, чем обычно бывает или чем ожидал наблюдатель. *Длинный* — это прилагательное, относящееся к большому полюсу длины. А *удлиненный* — это значение признака, сдвинутое по сравнению с нормой (серединой, нулем) в сторону большого полюса, но не достигшее его.

Таким образом, компаративная модель отпричастного прилагательного, как сам компаратив и как градативные глаголы, выражает идею относительно небольшого сдвига по шкале признака или параметра. Непроизводного прилагательного с таким значением нет, так что отпричастное прилагательное предоставляет говорящим недостающее средство для выражения этого значения.

КОЛИЧЕСТВЕННАЯ МОДЕЛЬ, или «ДОСТАТОЧНАЯ СТЕПЕНЬ» (*осведомленный источник*).

К этой группе относятся прилагательные с компонентом ‘много’. Именно компонент ‘много, значительное количество’ превращает причастие в качественное прилагательное (или относительное отпричастное прилагательное в качественное). Относительное прилагательное содержит отсылку, но не содержит идеи количества, ср. — *автомобильный завод* — ‘где производят автомобили’. При этом на заводе производят, конечно, большое количество автомобилей, но в относительное значение этот количественный смысл не входит; напротив, в случае *Москва — самый автомобильный город в нашей стране* благодаря компоненту ‘много [автомобилей]’ реализуется качественное значение.

Аналогично в случае отпричастных прилагательных:

насыщенный раствор — ‘содержит много растворенного вещества’;

осведомленный источник — ‘имеет много сведений’;

распространенное мнение — ‘его придерживаются многие люди’.

Качественное прилагательное всегда связанно с количественной шкалой, по которой «скользит» значение признака. Но обычные качественные прилагательные, скажем, параметрические, имеют довольно ограниченные «выразительные» возможности. Так, прилагательные «большого полюса» — *высокий*, *широкий*, *глубокий* — обозначают ‘больше нормы высоты, ширины, глубины’ (ср. [Sapir 1944], [Апресян 1974]). Но сам этот большой полюс имеет разные зоны, которые человеку хочется различить.

Например, *распространенное* (*мнение*) — это качественное прилагательное, т. к., в отличие от причастия (ср. *листовка, распространенная среди жильцов дома* — неизвестно — и не важно! — много ли жильцов), оно содержит идею относительно большого количества: *распространенное мнение* — это мнение достаточно большого количества людей (если бы его придерживалось мало людей, оно бы не называлось распространенным). Но это относительно большое количество все-таки отличается от очень большого количества, ср. *всеобщее мнение* или *переполненное метро*.

КАЧЕСТВЕННАЯ МОДЕЛЬ, или «СЛАБАЯ ОЦЕНКА» (*воспитанный юноша*).

Что входит в понятие *оснащенная лаборатория*? Человек оснащает лабораторию таким оборудованием, которое наилучшим образом приспособлено для реализации его целей. *Оснащенная лаборатория*, конечно, включает отсылку к событию *оснастить*, но семантический акцент делается на другом: она не просто содержит оборудование, но это оборудование оптимально для осуществления исследований. Сам по себе завоз оборудования (что тоже описывается глаголом *оснастить*) не делает лабораторию *оснащенной* (в качественном

значении), т. к. это может быть некомплектное или устаревшее оборудование. Т.е. признак *оснащенная* характеризует и качество оборудования, и вообще наличие условий для успешных исследований. Таким образом, *оснащенная* (*лаборатория*) значит ‘хорошо оснащенная’.

Аналогично:

продуманное решение не просто продумали, результатом этого процесса стал выбор наилучшего, оптимального варианта по сравнению с другими возможными решениями;

воспитанного юношу не просто воспитали (процесс воспитания, вообще говоря, не гарантирует выработку социально одобряемых качеств, хотя направлен именно на это). Качественное значение *воспитанный юноша* отражает не любой результат воспитания, а совершенно определенный — положительные качества, хорошее поведение (ср. *юноша*, *воспитанный в кадетском корпусе*, где причастие тоже отсылает к «интервалу воспитания», но ничего не сообщает о приобретенных качествах);

неустроенный быт не просто никто не устраивал, — он отличается отсутствием комфорта, неудобствами, т. е. имеет отрицательные характеристики как результат того, что его никто не устраивал;

запрещенный прием (в полемике) — это аморальный, неодобряемый прием, которым не следует пользоваться, как если бы его запретили. Показательно, что *запрещенный прием* формально никто не запрещал, т. е. для образования оценочного значения важно даже не реальное событие в прошлом (его может и не быть), а регулярное соотношение, естественная связь между событием и результатом: если объект имеет отрицательные качества, его следует запретить и обычно запрещают;

ср. также *подготовленные ребята* (‘хорошо подготовленные’, т. е. ‘получившие нужные навыки’, а не просто ‘прошедшие подготовку’); *приспособленное помещение* (‘хорошо приспособленное’), *наполненная жизнь* (‘интересная’), *заслуженный человек* (‘обладающий большими достоинствами’) и т. д.

Таким образом, у отпричастных прилагательных данной группы есть компонент ‘хорошо’ или ‘плохо’, поэтому данный тип значений можно назвать оценочным.

Иногда одному причастию могут соответствовать два разных адъективных значения –неоценочное и оценочное, — которые образуют своего рода пары:

неоценочное	vs.	оценочное
организованный туризм	—	*организованная работа, действия*;
неорганизованный туризм, отдых	—	*неорганизованные действия*;
согласованный митинг	—	*согласованные действия*;
запрещенные препараты	—	*запрещенный прием*.

Для оценочных значений важно не само глагольное событие (в некоторых случаях реального события нет, оно домысливается), а импликации из него. Причем эти импликации реализуют идею нормального хода событий (ср. использование понятия «нормального» положения вещей при описании

семантики противительных союзов, см., например, [Санников 1986; 1987]. В производном качественном значении фиксируется результат глагольного события. Теоретически результат может быть положительным, нулевым или отрицательным. Поскольку в данном случае глагольное событие — это целенаправленное действие человека, то естественным, ожидаемым, запланированным ходом событий считается достижение оптимального результата, в наибольшей степени соответствующего цели. Например, в случае *продуманного решения*, раз уж человек потратил время и усилия на обдумывание, естественным, ожидаемым (и желаемым) результатом является выработка хорошего решения. И наоборот, в случае *неустроенного быта*, — если человек не затрачивал усилий на устройство быта, естественным результатом будет наличие неудобств. Такова логика развития оценочных значений.

Если отпричастное прилагательное соотносится как с агентивным, так и с неагентивным глаголом, ср. *ограниченный*, *упрощенный* (*упростить* — *упроститься*), действует другая, тоже достаточно стандартная и распространенная импликация: ‘больше → лучше’, ‘меньше → хуже’, например:

ограниченный доступ — ограниченный ум;
упрощенная процедура — упрощенный подход.

ЗАМЕЧАНИЕ. Оценочные значения изредка развиваются и у активных отпричастных прилагательных, ср.: *знающий специалист*, *думающий студент*, *мыслящие люди*, *понимающие родители*, *любящий отец*. Иногда качественно-оценочное значение обнаруживается уже у производящего глагола, ср. *стоящий фильм — фильм стоит посмотреть*.

Очевидно, однако, что оценка в значении прилагательных данной группы занимает не такое место, как у прилагательных типа *прекрасный*, *чудесный*, *ужасный*, *чудовищный* и под., у которых лексическое значение целиком сводится к оценке. Признаки типа *подготовленные ребята*, *организованная работа*, *налаженный быт, ухоженные руки* и т. п. следует скорее отнести к разряду положительных качеств (или положительных характеристик), чем собственно оценок. Соответственно, *ограниченный ум*, *запрещенный прием*, *неустроенный быт* — негативные качества (характеристики).

Итак, мы можем констатировать, что процесс адъективации причастий является активным, а сама эта группа лексики постоянно пополняемой потому, что «поставляемые» этим процессом единицы отвечают номинативным запросам, номинативным потребностям говорящих. Хотя отпричастные прилагательные часто имеют семантические аналоги среди обычных глагольных и неглагольных прилагательных, они содержат важные семантические модели интерпретации признака, которых нет у других классов прилагательных и которые формируют недостающие типы характеристик объектов, недостающие классы обозначений. Наиболее многочисленные классы дают компаративная модель, а также модели «достаточной степени» и «слабой оценки».

Литература

1. *Апресян Ю. Д.* Лексическая семантика. Синонимические средства языка. М., 1974.
2. *Калакуцкая Л. П.* Адъективация причастий в современном русском литературном языке. М., 1971.
3. *Лопатин В. В.* Адъективация причастий в ее отношении к словообразованию // Вопросы языкознания. 1966, № 5. С. 37–48.
4. *Павлов В. М.* О разрядах имен прилагательных в русском языке // Вопросы языкознания. 1960, № 2. С. 65–70.
5. *Падучева Е. В.* Динамические модели в семантике лексики. М., 2004.
6. *Панов М. В.* Позиционная морфология русского языка. М., 1999.
7. *РР* 1973 — Русская разговорная речь. Отв. ред. Е. А. Земская. М., 1973.
8. *Санников В. З.* Значение союза но: Нарушение «нормального» положения вещей // Известия АН СССР. Сер. лит. и яз. 1986, № 5.
9. *Санников В. З.* Смысловой компонент ‘норма’ в значении русских сочинительных союзов // Вопросы кибернетики. М., 1987.
10. *Типология* 1983 — Типология результативных конструкций: результатив, статив, пассив, перфект / Под. ред. В. П. Недялкова. Л., 1983.
11. *Sapir E. Grading.* A study in semantics. // Philosophy of science, vol II, № 2. 1944.

RUSSIAN LEARNER PARALLEL CORPUS AS A TOOL FOR TRANSLATION STUDIES

Kutuzov A. B. (akutuzov72@gmail.com)
Lionbridge Technologies, Inc, Tampere, Finland

Kunilovskaya M. A. (mkunilovskaya@gmail.com),
Oschepkov A. Y. (aoschepkov@gmail.com),
Chepurkova A. Y. (AnnaChepurkova@yandex.ru)
Tyumen State University, Tyumen, Russia

The paper presents a project aimed at the development of a Russian Learner Parallel Corpus, discusses the existing analogues, describes the current status and the tasks in which it could be used. The existing parallel corpora contain (comparatively) "correct" translations; whereas the aim of the present project is to create a sufficiently large corpus of imperfectly translated Russian and English texts together with their sources and use it as a tool for translation studies, especially those related to translation mistakes. The new corpus will be a valuable resource for computational linguistics as it provides another way of getting data for evaluation which could be used to improve machine translation systems. As of now, the corpus is available online, it already contains nearly half a million word tokens and is growing. The main source of material is translations made by student translators in Russian universities.

Keywords: parallel corpora, corpus building, translation studies, translation mistakes, learner corpora

1. Introduction

The present paper is about the project of **Russian Learner Parallel Corpus (RLPC)**, which is currently under development by a group of Tyumen-based linguists. We will discuss the feasibility of such a corpus, similar projects that have been carried out elsewhere and describe the progress we have made in corpus building as well as the tasks which lie ahead and possible applications of this corpus.

2. Project outline and preceding research

Corpus linguistics has been enjoying the justified attention of translation studies for quite a long time, which is supported at least by the fact that there exist functioning translation and parallel corpora. The cooperation of corpus linguistics and translation studies gave birth to a wide range of translational corpora types. At the moment there are monolingual translational corpora (e. g., Translational English Corpus

by Mona Baker) and multilingual aligned corpora with texts created by professional translators (e. g., parallel sub-corpus of Russian National Corpus). The aims of such corpora to provide verified data on translational conformities account for the fact that they do not include amateur or non-professional translations a priori containing translational errors. Such translations are considered to be nuisance, a kind of noise undesirable in parallel corpora.

However, translational mistakes are of interest for researchers in the fields of translation studies and psycholinguists, as well as for methodologists and teachers directly involved with vocational translation training, as will be shown below. That's why the modern range of corpora certainly needs more multilingual translational corpora compiled from non-professional translations.

This is especially true for translations to and from Russian. It seems that there exist(ed) only one such corpus. Since 2004 the Department of Applied Linguistics of the Ulyanovsk State Technical University (Russia) carried out a similar project [Sosnina 2005]. Unfortunately, this corpus is not available on-line, which makes it impossible to make use of it, moreover, 1 million word tokens is not a sufficient number for universal deductions.

There is also the infamous MeLLANGE Learner Translator Corpus (http://corpus.leeds.ac.uk/mellange/ltc.html), the project completed in 2007 by a pull of European scientists and promoted by Department of Intercultural Studies and Applied Languages of the University Paris Diderot and the Société Française des Traducteurs (cf. Kübler 2008). This corpus is compiled of 440 student and professional translations, 350 words long on average, in at least 3 languages (English, German, French) covering 4 text types that were inter alia annotated for errors according to a customised error typology (http://corpus.leeds.ac.uk/mellange/images/mellange_error_typology_en.jpg). The project has a developed query interface which is compatible with larger parallel corpora developed at University of Leeds Centre for Translation Studies. Though this project is a powerful tool for creating training materials and educating translators, it does not include texts in Russian and is quite limited in terms of text types and overall scope.

Thus, there is a certain vacuum, which our project is supposed to fill.

3. Aims of the project

Within the current project we aim to design and compile **Russian Learner Parallel Corpus** which will include non-professional (student) translations, aligned with their source texts. It would be large enough to give reliable data on translation mistakes and easily-available via the Internet to be used in translation studies research, to create training materials and to inform translation pedagogy. The target corpus size is 10 million word tokens. It will include source and translated texts in two directions: “En-Ru” and “Ru-En”. Certainly, there can (and do) exist multiple translation variants of the same source.

The Corpus is compiled of translations made by Translation Studies university students with various levels of translation experience from several Russian universities.

Their translations do regularly contain mistakes, they already exist in digital form and they are comparatively easy to obtain.

It should be stressed that our Corpus is after translation mistakes. Though it is intuitively clear what sort of imperfections can a learner translation carry we will offer a few examples that are in no way meant to be exhaustive or limitative or even representative. They are included here merely for the sake of illustration.

Source 1 (from an essay on car industry):

(1) *"The problem with the car, a century on, is that we have allowed Henry Ford and not Henry Royce (of Rolls-Royce fame) to be the winner. For the most part, the car is now an ugly, commonplace form of transport rather than a beautiful toy or exquisite work of engineering. Mass-market cars may be democratic, but with a few shining exceptions (the Mini; the Fiat 500; the Volkswagen 'Beetle') — the vast majority are about as inspiring as a dishwasher".*

Translation 1.1.

(2) *"Сейчас, век спустя, проблема автомобилей в том, что мы позволили Генри Форду, а не Генри Ройсу (основатель компании Ролс-Ройс) стать победителем. Автомобиль сейчас, по большей части, — просто средство передвижения, а не красивая игрушка или привлекающее внимание произведение инженерного искусства. Автомобили массового производства, за редким исключением, весьма неплохи. Но некоторые представители этого класса (например, Мини, Фиат 500, Фольтсваген «Жук») вызывают не больше эмоций, чем посудомоечная машина".*

It seems that the message in the Russian variant of the first sentence is unclear, the implication of "winning" require some explanation such as the one offered in the version below. Besides the overall wording of the sentence in 1.1 is inelegant and clumsy which is especially obvious in the collocation "проблема автомобилей". A much more grievous mistake can be found in sentence 3. The message in translation is opposite to that in the source (cf. translation 1.2 below) and contradicts the logic of the sentence immediately above which contributes to the general target text incongruity. The situation is further aggravated by the extension this mistake has in the closing sentence. This type of translation error is usually referred to as content mistake and is opposed to language (or form) mistakes such as inappropriate collocation, wrong lexical choice (ex. «игрушка» in this context) or grammar mistake.

Translation 1.2.

(3) *"Но дело в том, что сегодня у нас были бы совершенно другие машины, если бы сто лет назад в конкурентной борьбе победил не Генри Форд, а Генри Ройс (основатель марки Rolls-Royce). По большей части, современный автомобиль — это заурядное средство передвижения, а не изысканное*

инженерное творение. Конечно, автомобили массового производства доступны, но, за исключением ярких примеров, таких как Mini, Fiat, Volkswagen (Жук), в большинстве своем, они вызывают не больше эмоций, чем посудомоечная машина".

Source 2 (from a company profile description):

(4) *"Развитие малых предприятий тормозит недостаток оборотных средств".*

Translation 2.1. *"Development of small businesses prevents lack of circulating assets".*

A typical mistake for trainee translators working into a foreign language is copying the syntax of the mother-tongue source as is the case in the example above. Here a form mistake interferes with the content. There is also a terminology mistake in rendering "оборотные средства" which are usually referred to in English as "current assets". Below there are two more example of glaring literalisms in translation: a syntactic and a lexical ones respectively:

Source 3:

(5) *"Предпринимателю необходимо в вежливой форме выяснить, на каком основании производится проверка".*

Translation 3.1.

(6) *"To the businessman it is necessary in polite form to find out, on what basis check is made".*

Source 4:

(7) *"Юридические лица в пределах своей компетенции"*

Translation 4.1.

(8) *"Legal persons"*; *"inside its competence".*

4. What one can do with this corpus?

The study of translated texts versus sources yields valuable results for computer-powered monolingual and contrastive research. A new scope to such studies can be added by the analysis of *'non-professional' 'imperfect'* translations based on the adequate and representational corpus which our project is aimed at.

Below we will deliberately focus on the most workable applications of the Russian Learner Parallel Corpus among many more other possible uses. They seem to fall into three major categories:

1) Translation Studies Research in translation mistakes

The analysis of a variety of imperfect translations against originals can be used to see which of the mistakes are translation-wise relevant and which can be dismissed as formal linguistic faults having no detrimental effect on the translation in general. With any luck this analysis can be formalized on the basis of frequencies in translation memory-like systems.

For example, mistakes related to verb aspect choice (or noun case, or subject and predicate agreement) in Russian translations usually do not corrupt semantic equivalence to the source. At the same time, lexical mistakes in translating connotations or metaphors lead to text tonality change and distortion of text pragmatics. Similarly modelling syntax after foreign patterns often damages semantics of the text.

Further on, generalizations on the translational methods applied in erroneous units can show which of the strategies can be considered inapplicable in particular text-types due to the distortions in pragmatic and semantic plains of the source text and what sort of mistakes more often cause distortions of the source textual and interpersonal meta-functions.

The analysis of imperfect translations can be used to discover translational universals and conformities or translational units that pose no problems in a certain language combination in a certain text-type.

2) Translator Training Methodology

Even wider and immediately practicable applications of the Corpus (which so far lacks translation mistakes mark-up and is basically a collection of student translations) can be seen within translator training sphere. This collection can be used as material in post-editing classes and in the translation studies courses where students practice to spot mistakes, explain their nature and offer other variants as well as to compare different ways of expressing the same idea. It also provides ample material to develop learners insight into the much-discussed issue of the quality in translation. It helps to teach the difference between standard and individual translational problems and generalize about ways of overcoming them. On the basis of the Corpus a methodologist can also make didactic conclusions and target mistakes typical for Russian translation students in their training, including target language mistakes induced by source texts and revealed through the comparative analysis of the translation subcorpus and original target language texts.

3) Computational Linguistics

The Corpus provides the material for comparative analysis of mistakes typical for machine and human translations and draw conclusions on the differences in the processes involved.

The machine-translation end of this Corpus applications also involves the possible use of the “negative” material (generalizations about wrong translations of certain language units for the given language combination) in developing software. Comparative frequencies of corresponding words in two languages and in translational

texts can also give useful insights about evaluating and improving machine translation systems.

The Corpus can be used to test or develop algorithms for evaluating the quality of text, such as BLEU (Bilingual Evaluation Understudy) [Papineni 2002].

And finally, the accessibility of the Corpus via the Internet implies the possibility of verification of the findings based on it and also the possibility of offering other explanations by appealing to parameters that may have been downplayed or ignored in previous studies [Baker, 2004].

5. Quantitative and qualitative data about the Corpus

As of now the collected corpus (available on-line in raw form — http://tc.utmn.ru/files/trc.zip) consists of 1146 texts, English and Russian, sources and corresponding translations. The accurate numbers are given in the table below.

Texts

	Source	Translation	Total
English	100	513	613
Russian	25	508	533
Total	125	1021	1146

These texts contain **467 513 word tokens, 51 % of them being English and 49 percent Russian**. We continue to add texts and will do so until we reach 1 million tokens. This corpus will be made available on-line with query and feedback tools. After that we plan to analyse all the feedback we will have to this moment from our users: researchers of translation errors and conformities and teachers using the "negative" data from the Corpus to create translator training materials. It is especially important for use to receive comments about corpus structure, meta-data, GUI options and actual cases of corpus usage.

After making adjustments to the Corpus structure if necessary we plan to continue expanding the Corpus to reach 10 million word tokens. We are going to set up an on-line contribution option for volunteers to submit source and translated texts, thus employing the power of crowd-sourcing. However, we will still strive to preserve the present ratio of English and Russian translations.

At present we collected texts from Tyumen State University (about 900 texts), Moscow State University (about 70 texts) and Udmurt State University (about 100 texts). Texts from Chelyabinsk State University and Nizhny Novgorod State Linguistic University are still being processed.

Technically, the Corpus is organized in plain text files and header files, according to standards of Translational English Corpus (http://www.llc.manchester.ac.uk/ctis/research/english-corpus). Header files contain student and text meta data for corresponding text files which is arranged in the following fields which present information that can actually be found in the Corpus:

1. Translator' sex
2. University level (Year of study)
3. Assessed translation (Mark received for this translation)
4. Draft or final version
5. Genre of the text (Academic/News(informational)/Essay/Interview/Tech/Fictional/Educational/Encyclopaedia/Speech/Business Letters
6. Type of translation (exam or routine work)
7. Situation of translation (home or classroom)
8. Translation date (by year)
9. Corpus subset (whether the text is a source or a translation)
10. University which has submitted texts

This annotation does not include translation mistakes mark-up which we only plan to develop and introduce at some later stage. Such mark-up, however, will be available for a limited subset within the Corpus because it will have to be made by hand (see below for details).

We plan to use TEC Browser tool (http://modnlp.berlios.de) in order to allow users to query particular sub-corpora. For example, one could look for draft translations made by male students of the Tyumen state university in 2009 and compare them with translations made by female students.

This web application (based on Modnlp libraries) will allow to query the corpus without downloading and installing any software. Queries in bilingual corpus are organized in 3 stages:

1. Looking up text string as per user query.
2. Strings found during the first stage are checked for their translations with the help of bilingual TMX file.
3. All the translations found during the second stage are checked for their equivalents in the other language.

The result of a query is the text string mentioned in the query and all variants of its translations.

Texts in the corpus will be POS-tagged and syntactically marked-up, possibly with the help of Mystem and Synan respectively.

6. Further research

At present the Corpus is still under development and within the next few months we will have to pass the following milestones to reach our objective of providing a comprehensive tool to translator trainers and translation studies academia:

1. Increase the Corpus scope while keeping its structure intact.
2. Develop graphical query interface available on-line. It would be based on TEC Browser client, but we need to take considerable steps to expand its functionality (presently it doesn't work with aligned texts).

3. Provide morphological and syntactic mark-up to make it possible to use the corpus in grammatical contrastive and translation research.
4. It is ambitious but not altogether impossible to supply a smaller sub-corpus with descriptive linguistic mark-up employing XML. Such mark-up can be based on the most basic and universally recognized types of mistakes such as form mistakes and content mistakes discriminated on the degree they interfere with rendering the source message.
5. One can go on and create a GUI which will allow a reviser (such as a teacher or a critic) to mark mistakes electronically and therefore make a revised copy of a student's translation easily and automatically assessable and a ready-made material for further extension of the Russian Learner Parallel Corpus into the bargain.

References

1. *Baker M.* (2004), A corpus-based view of similarity and difference in translation, International Journal of Corpus Linguistics, Volume 9, Number 2, 2004, pp. 167–193
2. *Kübler N.* A Comparable Learner Translator Corpus: creation and use. Proceedings of the Comparable Corpora Workshop of the LREC Conference, May 31 2008, Marrakech, Maroc, pp 73–78
3. *Papineni K., Roukos S., Ward T., Zhu W.* BLEU: a method for automatic evaluation of machine translation. ACL 02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics Association for Computational Linguistics. Stroudsburg, PA, USA, 2002, pp 311–318
4. *Sosnina E.* Russian Translation Learner Corpus: The First Insights. The proceedings of the 6 international scientific conference «Interactive systems: problems of human-computer interaction»,Ulyanovsk: UlSTU, 2005

К ПОСТРОЕНИЮ ИНВЕНТАРЯ РУССКИХ ИМЕННЫХ КОНСТРУКЦИЙ[1]

Ляшевская О. Н. (olesar@gmail.com)
Национальный исследовательский университет
Высшая школа экономики, Москва, Россия

Митрофанова О. А. (alkonost-om@yandex.ru)
Санкт-Петербургский государственный университет,
Санкт-Петербург, Россия

Грачкова М. А. (maaag86@mail.ru)
Санкт-Петербургский государственный университет,
Санкт-Петербург, Россия

Романов С. В. (complefor@rambler.ru)
ЗАО «Интернет-Проекты», Санкт-Петербург, Россия

Шиморина А. С. (shinas@yandex.ru)
Институт лингвистических исследований РАН,
Санкт-Петербург, Россия

Шурыгина А. С. (sanyana@gmail.com)
Российский государственный гуманитарный университет,
Москва, Россия

В докладе обсуждаются результаты экспериментов по автоматическому выделению конструкций, проводимых на материале Национального корпуса русского языка (НКРЯ). С этой целью разработан компьютерный инструмент, позволяющий извлекать и обрабатывать сочетаемостные данные из выборок НКРЯ. В качестве целевых слов выступают русские имена существительные. Для каждого из целевых слов получены списки конструкций — наиболее частотные сочетания, включающие целевое слово, частотные лексико-семантические теги — контекстные маркеры того или иного значения целевого слова, а также частотные леммы, характеризуемые этими семантическими тегами. Например: ВИД (*разновидность, тип*) + **r:abstr t:sport:** *спорт, футбол, биатлон* и т. д. Выделенные конструкции систематизируются по структуре и лексико-семантическому наполнению. В заключение проводится проверка результатов экспериментов, предполагающая сравнение списков конструкций со сведениями о коллокациях, устойчивых сочетаниях и т. д., зарегистрированных в различных лингвистических источниках (сервисы поиска биграмм, словари).

Ключевые слова: конструкции, корпус текстов, лексико-семантическая разметка, существительные, русский язык

---

[1] Работа выполнена при финансовой поддержке РФФИ (проект 10-06-00586-а) и программы фундаментальных исследований Президиума РАН «Корпусная лингвистика» (проект FrameBank), а также проекта НИР «Модель интегрированного программно-лингвистического комплекса для создания специализированных корпусов русского языка».

1. Введение

Создание лексикографических ресурсов на основе корпусов текстов — одно из продуктивных направлений современных исследований в области автоматической лексикографии и компьютерной лингвистики. Главное достоинство корпуса текстов как источника лингвистических данных заключается в том, что на его основе возможно создание не только традиционных ресурсов (в частности, словарных баз данных), но и ресурсов, объектом описания в которых являются довольно сложные лингвистические явления, в частности, конструкции (см., например, **http://dict.ruslang.ru/**, Kustova 2011). Так, автоматическое построение каталога русских конструкций на базе Национального корпуса русского языка (НКРЯ, **http://www.ruscorpora.ru/**) — цель нашего проекта, реализуемого коллективами НКРЯ и кафедры математической лингвистики СПбГУ.

В исследованиях, опирающихся на конструкционные подходы в лингвистике, используется широкое понимание конструкции: в общем случае, это сложный знак, значение которого не выводится из значения составляющих и компоненты которого взаимодействуют между собой и взаимообуславливают друг друга (Rakhilina 2010; Oskol'skaya, Say 2010, Ovsjannikova et al. 2011). Лексически-ориентированные конструкции предполагают организацию вокруг одного или нескольких фиксированных лексических элементов (например, коллокация, коллострукция, морфо-синтаксическая схема, модель управления, комбинация с лексической функцией, синтаксическая группа с фиксированными лексическими слотами, или конструкция «малого синтаксиса», и др.).

Идентификация лексических конструкций в корпусе связана с двумя важными задачами. Во-первых, сочетаемость и синтаксис довольно больших слоев лексики (прежде всего, имен существительных, прилагательных и наречий) описана в русистике совершенно недостаточно (из положительных примеров можно назвать TKS 1984, Apresjan 2003, 2010 и нек. др., охватывающие лишь отдельные единицы). Во-вторых, процедура выделения конструкций связана с разрешением лексико-семантической неоднозначности (Mitrofanova et al. 2008; Mitrofanova, Lyashevskaya 2009; Lyashevskaya et al. 2011; Shimorina, Grachkova 2011). Если слово многозначно, то, в идеале, каждое значение характеризуется своим кругом конструкций — тем самым, анализ семейств близких конструкций дает информацию, позволяющую распознавать и разграничивать значения многозначного слова. Кроме того, эксперимент с автоматическим выделением конструкций позволяет проверить теоретическую гипотезу о генерализациях в ходе усвоения языка ребенком (Dąbrowska 2004). Эта гипотеза гласит, что человеку свойственно собобщать сходные часто повторяемые цепочки слов в виде единиц более абстрактного уровня (=конструкций), причем ключами для генерализации могут служить единицы морфологического, лексемного, синтаксического и любого другого уровня.

Таким образом, под конструкцией в нашем проекте понимается сочетание целевого слова и контекстных маркеров его значения, характеризующееся частотностью и устойчивостью. Данная трактовка конструкций согласуется с основными идеями Грамматики конструкций (Fillmore 1988; Goldberg 1995,

2006; Tomasello 2003; Kuznetsova 2007). Как контекстные маркеры рассматриваются теги, доступные в многоуровневой разметке контекстов НКРЯ: теги лемм (*lex* — лексема, которой принадлежит словоформа), морфологические теги (*gr* — грамматические признаки словоформ: частеречная принадлежность, значения грамматических категорий и т. д.), лексико-семантические теги (*sem* — признаки, указывающие на принадлежность слова к определенному лексико-семантическому классу[2]).

Примерами конструкций могут служить следующие сочетания существительного *вид* с левосторонними/правосторонними соседями, маркирующими различные его значения:

(1) значение **r:abstr t:perc der:v**: *внешность, видимый облик; состояние*
внешний + ВИД

ВИД + **r:concr t:hum:** *человек, женщина, король, друг, сосед, господин, товарищ* и т. д.

(2) значение **r:abstr r:concr pt:set sc:X**: *разновидность, тип*

r:card:pauc: *два, три, четыре, оба* и т. д. + ВИД

r:abstr der:v: *использование, стоимость, смешение, доминирование, сохранение* и т. д. + ВИД

ВИД + **r:abstr t:sport:** *спорт, футбол, физкультура, картинг, биатлон* и т. д.

и др. В качестве более сложных конструкций могут выступать такие цепочки, как *представить* + Sacc + *В* + *ВИДЕ* + Sgen.

В настоящее время ведется активное обсуждение методов и алгоритмов выделения конструкций (Sahlgren, Knutsson 2009; Proceedings of the NAACL… 2010; Wible, Tsao 2010). Большие успехи достигнуты в области извлечения *n*-грамм (коллокаций, неоднословных целостностей — ср. (Manning, Schütze 2002, среди многих других; для русского языка –Jagunova, Pivovarova 2011). Однако идиоматизированные конструкции, а также конструкции с нестандартной синтаксической структурой, хотя они подробно описаны в исследовательской литературе (Borisova 1995, Iordanskaya, Mel'chuk 2007), представляют серьезную проблему в автоматической обработке текста. Автоматизация выделения конструкций из русскоязычных текстов усложняется такими особенностями материала, как свободный порядок слов и богатое словоизменение, снижающие предсказуемость формальной организации конструкций. По этой причине выделение конструкций с опорой на *n*-граммы, допустимое, например, в англоязычных текстах, в нашем случае не всегда приводит к желаемым результатам:

---

[2] Например, **r:concr** — предметные имена (*девочка, стол, молоко*); **r:abstr** — непредметные имена (*вождение, яркость, время*); **t:hum** — лица (*человек, учитель*); **t:stuff** — вещества и материалы (*вода, песок, тесто, жесть, шелк*); **t:constr —** здания и сооружения (*дом, шалаш, мост*); **t:tool** — инструменты и приспособления (*молоток, палка, пуговица, машина*); **t:move** — движение (*беготня, вынос, качка*); **t:perc** — восприятие (*осязание, слух, видимость, взгляд, зрелище*); **t:space** — пространство и место (*космос, город, тайга, овраг, вход*); **t:psych** — психическая сфера (*апатия, безумие, вдохновение, спокойствие*); **t:speech** — речь (*дискуссия, молва, ахинея, реплика, подковырка*); **r:qual** — качественные (*хороший, большой*); **r:rel** — относительные (*деревянный, лунный*); и т. д., подробнее см.: **http://www.ruscorpora.ru/corpora-sem.html.**

фрагменты конструкций могут не иметь явного выражения в рамках заданного контекста, могут выходить за пределы синтаксических групп и т. д.

Существует ряд проектов, в которых особое внимание уделяется формализации лексико-синтаксических связей единиц текста, среди них есть исследования и на материале русского языка: Word Sketches для русского языка (Zakharov, Khohlova 2010), работы по извлечению лексико-синтаксических шаблонов (Bolshakova et al. 2007), по автоматическому построению словарей сочетаемости (Gel'buh et al. 2004). Тем не менее, электронные ресурсы, отражающие сочетаемостные предпочтения и рамки валентностей русской лексики, недостаточно разработаны. Если бы в распоряжении исследователей были такие ресурсы, как FrameNet или PropBank для русского языка, то это существенно облегчило задачу каталогизации конструкций (Lyashevskaya, Kuznetsova 2009).

В данной статье обсуждается один из важнейших аспектов проекта, а именно, построение шаблонов именных конструкций на основе контекстных маркеров значений целевых слов.

2. Лингвистические данные

Источником данных об именных конструкциях служат выборки контекстов из НКРЯ, крупнейшего русскоязычного корпуса с многоярусной разметкой. В центре внимания исследовательской группы находятся (1) отдельные русские существительные, представляющие различные лексико-семантические группы: *дом*, *вид*, *орган, лук, глава* и т. д., (2) целостные лексико-семантические группы существительных — названий инструментов (*бритва, веник, весло, карандаш, коса, лом, лопата, метла, ножницы, топор, щетка*), обозначений речевых действий (*дискуссия, комплимент, обращение, обсуждение, ответ, похвала, рекомендация, вопрос, вранье, выражение*) и т. д. Анализируемые лексемы отличаются количеством значений, характером развития полисемии/омонимии, степенью связанности значений между собой. В нашем исследовании производится условное приравнивание омонимичных коррелятов к многозначным словам (Rakhilina et al. 2006). Разметка значений слов в контекстах НКРЯ проводилась на основе Семантического словаря НКРЯ. В табл. 1 приводится филиация значений слова *глава* с указанием их лексико-семантической аннотации и частотности в выборке, в табл. 2 — пример разметки контекстов со словом *глава*.

Таблица 1. Филиация значений слов *глава*

Значения	**Лексико-семантическая аннотация**	**Примеры**	**Число контекстов: 1056, из них**
m1. То же, что голова	**r:concr pt:partb pc:hum** предметное имя, часть тела человека	*Склонить главу.*	8

Значения	Лексико-семантическая аннотация	Примеры	Число контекстов: 1056, из них
m2. ***Во главе*** *— возглавляя кого-что-н., впереди кого-чего-н.* ***Во главу угла ставить*** *— считать самым важным*	**r:concr der:shift dt:partb** предметное имя, метонимический перенос	*Идти во главе колонны.*	136
m3. Купол церкви	**r:concr pt:part pc:constr** предметное имя, часть здания	*Главы собора.*	1
m4. Руководитель, начальник, старший по положению	**r:concr t:hum** предметное имя, лицо	*Глава государства.*	299
m5. Раздел книги, статьи	**r:concr t:text pt:part pc:text** предметное имя, текст, часть текста	*Глава книги.*	612

Таблица 2. Пример разметки контекстов со словом *глава*

Значение	*m5. Раздел книги, статьи*
Контекст	*Глава 8, в которой родители получают странные письма.*
Разметка	<w><ana lex='глава' gr='S,f,inan=sg,nom' SEMF='r:concr t:text pt:part pc:text '/></ana>Глав`а</w> <w><ana lex='8' gr='NUM=ciph'/></ana>8</w>, <w><ana lex='в' gr='PR'/></ana>в</w> <w><ana lex='который' gr='A-PRO=f,sg,loc' sem='r:rel | r:rel t:ord '/></ana>кот`орой</w> <w><ana lex='родитель' gr='S,m,anim=pl,nom' sem='r:concr t:hum:kin d:nag der:v '/></ana>род`ители</w> <w><ana lex='получать' gr='V,ipf,tran=pl,act,praes,3p,indic,act' sem='d:pref der:v '/></ana>получ`ают</w> <w><ana lex='странный' gr='A=pl,acc,plen' sem='r:qual '/></ana>стр`анные</w> <w><ana lex='письмо' gr='S,n,inan=pl,acc' sem='r:concr t:text'/></ana>п`исьма</w>

3. Компьютерный инструмент выделения конструкций

Компьютерный инструмент выделения конструкций позволяет выполнять автоматическую классификацию контекстов, с этой целью создается векторная модель экспериментальной выборки; в качестве базового алгоритма выбран

алгоритм классификации с учителем. Данные процедуры осуществляются с помощью программного обеспечения, разрабатываемого С. В. Романовым на языке Python. Инструмент работает в двух режимах: формирование классов контекстов, соотносимых с отдельными значениями целевого слова; генерация списков кандидатов в конструкции — наиболее частотных сочетаний, в которых реализуется то или иное значение целевого слова. Автоматическое выделение конструкций производится на основе статистических данных о сочетаемости целевых слов и левосторонних/правосторонних соседей — контекстных маркеров их значений: тегов *lex*, *gr* и *sem*. Сочетаемостная информация извлекается из обучающей выборки. Возможно варьирование таких параметров экспериментов, как ширина контекстного окна [*-l*; *+r*], обработка с учетом/без учета весов контекстных элементов. Результат работы программы отражается в виде списка частотных комбинаций целевого слова и статистически значимых контекстных маркеров, с указанием данных о частоте встречаемости этих сочетаний и с перечнями лексем, реализующих значения контекстных маркеров в их составе. Пример выдачи программы приведен в табл. 3. Подобные списки подлежат постредактированию, в ходе которого отсеиваются сочетания, малоинформативные с точки зрения определения значения целевого слова в контексте (например, для целевого слова *ножницы* таковыми признаны сочетания с местоимениями).

Таблица 3. Пример выдачи списка кандидатов в конструкции для слова ***НОЖНИЦЫ***

Параметры обработки данных	Лемма: *ножницы* Значение: **r:concr t:tool:instr der:s** (*режущий инструмент из двух раздвигающихся лезвий с кольцеобразными ручками*) Контекстное окно: [-1; +1] Объем выборки: 683 контекстов
Кандидаты в конструкции с левосторонними соседями	**t:impact ca:caus d:root:** резать(10) кромсать(1) стричь(4) **t:poss ca:caus:** взять(17) **r:dem:** этот(11) такой(1) тот(1) **t:impact ca:caus d:pref der:v:** вырезать(9) обрезать(4) надрезать(1) обстригать(1) **r:rel der:s dt:space:** садовый(19) …
Кандидаты в конструкции с правосторонними соседями	**r:poss:** ее(3) свой(1) наш(2) мой(2) **r:rel | r:rel t:ord:** который(8) **r:spec:** каждый(1) сам(1) весь(5) **t:impact ca:caus d:pref:** нарезать(2) срезать(1) перерезать(3) искромсать(1) **t:impact ca:caus d:pref der:v:** вырезать(4) надрезать(3) обрезать(1) **r:rel:** что(3) известный(1) дикий(1) лишний(1) обыкновенный(1) **r:pers:** она(2) я(3) он(1) они(2) мы(1) вы(1) …

4. Предварительные результаты экспериментов по выделению конструкций

Эксперименты по выделению конструкций проводятся в несколько этапов. Сначала для каждого значения рассматриваемых целевых слов составляется список контекстов его употребления, далее из контекстов автоматически извлекается наиболее частотная лексико-семантическая и морфологическая информация о контекстных маркерах значения в заданном окне. Выделение конструкций изначально производится в пределах узкого контекстного окна [-1; +1], где высока вероятность встретить контактные контекстные элементы, входящие в устойчивые словосочетания с исследуемыми словами. Рассматриваются также и другие параметры контекстного окна, важные для диагностики дистантных контекстных маркеров и выявления конструкций со структурой, более сложной по сравнению с биграммами, в связи с этим в проводимых экспериментах окно расширяется до [-3;+4]. Далее формируются морфологические модели конструкций, описывается их лексико-семантическое наполнение. Методика лингвистического анализа морфологических моделей конструкций и их лексико-семантического наполнения, использованная в нашем исследовании, основывается на опыте анализа сочетаемостных предпочтений слов разных частей речи (Mitrofanova, Belik, Kadina 2008).

Очевидно, что отношения между элементами конструкций не сводимы к аддитивным. В ряде случаев конструкции реализуют разнообразные лексические функции (Mel'chuk 1999/1974, 2007), как, например, конструкции, выделенные для слова *комплимент* из лексико-семантической группы речи (значение **t:speech r:abstr**: *любезные, приятные слова, лестный отзыв*):

Mult (название типовой «частичной» совокупности): **pt:set r:concr:** *куча* + комплимент

Ver (правильный, соответствующий, какой следует): **ev r:rel:** *уместный* + комплимент

AntiVer **r:rel:** *сомнительный* + комплимент

$Oper_1$ (глагол, связывающий название первого актанта в роли подлежащего с названием ситуации в роли первого дополнения): **d:pref der:v:** *выражать* + комплимент; **d:root:** *делать* + комплимент.

Такие конструкции были выявлены в выборках из НКРЯ и зафиксированы в ТКС (TKS 1984).

При анализе контекстных данных, полученных для инструментальных существительных, мы наблюдаем разнообразие конструкций с точки зрения сложности (двучленные, трехчленные и т. д.), степени устойчивости, морфосинтаксических свойств. Выявленные конструкции допускают обобщение до шаблонов, реализующих инструментальное значение.

Чаще всего встречаются двучленные конструкции — свободные и устойчивые сочетания вида *A+S / S+A, V+S / S+V, S+S* с различным лексико-семантическим наполнением.

(1) *A+S / S+A*: в таких сочетаниях выражаются
- разнообразные свойства инструмента: **r:rel ev d:neg der:a:** *безопасный* + БРИТВА; **r:rel ev:** *опасный* + БРИТВА; **r:qual t:physq ev:** *острый* + БРИТВА; **r:qual t:physst:** *сухой, мокрый* + ВЕНИК; **r:qual t:physq:** *мягкий, жесткий, гибкий, твердый* + ЩЕТКА; **r:qual t:physq:color der:s dt:color dt:abstr:** *цветной* + КАРАНДАШ и т. д.;
- тип инструмента: **r:rel der:s:** *механический, электрический* + БРИТВА; **r:rel der:s:** *химический* + КАРАНДАШ; **r:rel der:s dt:tool:instr dt:tool:** *совковый* + ЛОПАТА и т. д.;
- субъекты, использующие инструменты: **r:rel der:s dt:hum:** *саперный, дворницкий* + ЛОПАТА и т. д.;
- сфера применения инструмента: **r:rel der:s dt:space:** *садовый* + НОЖНИЦЫ; **r:rel der:s:** *маникюрный, хирургический* + НОЖНИЦЫ; **r:rel der:s dt:tool:cloth dt:tool:** *платяной, сапожный, обувной* + ЩЕТКА и т. д.;
- материал, из которого изготовлен инструмент: **r:rel der:s dt:stuff:** *деревянный, фанерный, железный, пластмассовый, металлический* + ЛОПАТА; **r:rel der:s dt:stuff:** *металлический, пластиковый, железный, пластмассовый, капроновый* + ЩЕТКА; *S+A*: ЩЕТКА + **r:rel der:s dt:stuff:** *пенный, абразивный* и т. д.;
- размер инструмента: **r:qual t:size:max:** *широкий, огромный, большой* + ЛОПАТА и т. д.;
- форма инструмента: **r:qual t:physq:form:** *круглый, плоский* + ЩЕТКА и т. д.

(2) *V+S / S+V*: в таких сочетаниях выражаются
- действия, производимые при помощи инструмента: БРИТВА + **t:impact ca:caus d:pref:** *срезать*; БРИТВА+ **t:impact ca:caus d:pref | t:impact ca:caus d:pref**: *разрезать*; ВЕНИК + **d:pref der:v | t:move ca:caus d:pref der:v:** *подметать*; **t:move d:pref der:v:** *огребать, подгребать* + ВЕСЛО; **d:impf d:pref der:v:** *исписывать, подчеркивать* + КАРАНДАШ; **d:pref der:v:** *нарисовать, отмечать* + КАРАНДАШ; КАРАНДАШ + **d:root:** *писать*; КАРАНДАШ + **d:pref der:v:** *помечать, нарисовать, изображать*; КАРАНДАШ + **d:impf d:pref der:v:** *подчеркивать, записывать*; ЛОПАТА + **d:impf d:pref der:v:** *откапывать, разравнивать, расковыривать, раскидывать, перемешивать, перекапывать*; **t:move ca:caus d:root:** *махать, гонять, мести* + МЕТЛА; МЕТЛА + **t:move ca:caus d:root:** *мести, гонять*; МЕТЛА + **t:move d:pref der:v:** *сметать, выгонять, сгребать, отгонять*; **t:impact ca:caus d:root**: *резать, стричь, кромсать, ранить* + НОЖНИЦЫ; **t:impact ca:caus d:pref der:v:** *вырезать, обрезать, обстригать, надрезать* + НОЖНИЦЫ; НОЖНИЦЫ + **t:impact ca:caus d:pref:** *перерезать, нарезать, срезать, взрезать, откромсать, искромсать, сломать, прорезать*; НОЖНИЦЫ + **t:impact ca:caus d:pref der:v:** *обрезать, вырезать, надрезать*; НОЖНИЦЫ +**t:impact ca:caus d:root:** *кромсать, резать, стричь*; **t:poss ca:caus:** *взять* + НОЖНИЦЫ; **t:impact d:pref der:v:** *вырубать, отрубать, приглушать, разрубать,*

обрубать, врубать, изрубать + ТОПОР; ТОПОР + **t:impact ca:caus d:root:** *бить, рубить, тесать*; ТОПОР + **t:impact d:pref der:v:** *разрубать, отрубать, обрубать, скреплять, прорубать, вырубать*; **t:poss ca:caus:** *взять* + ТОПОР и т. д.

– действия, производимые применительно к инструменту: ВЕНИК + **d:impf d:pref der:v:** *ошпаривать, раскидывать, наготавливать, заготавливать, пересушивать*; весло + **d:root:** *бросать*; МЕТЛА + **t:impact:creat d:root:** *вязать* и т. д.

(3) *S* + *S*: в таких сочетаниях выражаются

– совместно используемые инструменты: ВЕНИК + **r:concr t:tool:dish top:contain:** *ведро*; МЕТЛА + **r:concr t:tool:instr:** *лопата*, *швабра*, *щетка*, *веник*; ЛОПАТА + **r:concr t:tool:instr der:v:** *метла, грабли, скребок, тяпка*; ЩЕТКА + **r:concr t:tool:instr der:v:** *расческа, шило, чесалка, скребница, скребок, метла, бритва*; и т. д.

– вещества, материалы, задействованные при использовании инструмента: БРИТВА + **r:concr t:stuff:** *одеколон*; КАРАНДАШ + **r:concr t:stuff:** *ватман, бумага, ластик, акварель*; **r:concr t:stuff:** *уголь, бумага, пастель, ластик* + КАРАНДАШ; ЛОПАТА + **r:concr t:stuff:** *глина, навоз, цемент, уголь, грунт, песок, бетон*; ЩЕТКА + **r:concr t:stuff t:tool pt:qtm qc:stuff der:v:** *мыло*; **r:concr t:stuff:** *паста, гуталин, шампунь, песок, алюминий, вакса, вельвет, бархотка, хрусталь, наст, порошок, одеколон, стекло, ковер, сода,* + ЩЕТКА и т. д.

– субъекты, использующие инструмент: **r:concr t:hum der:s:** *сменщик, девушка, дружинник, сибиряк, мужик* + ВЕСЛО; МЕТЛА + **r:concr t:hum t:prof der:s:** *дворник*; **r:concr t:hum t:prof der:s:** *дворник* + МЕТЛА; ТОПОР + **r:concr t:hum:** *вельможа, человек, крестьянин, сэр, друг, дитя* и т. д.

– объекты, частью которых является инструмент: ВЕСЛО + **r:concr t:tool:transp:** *лодка, ушкуй, байдарка, ладья, крейсер*; **r:concr t:tool:transp:** *галера, корабль, лодка* + ВЕСЛО и т. д.

– части тела человека: БРИТВА + **r:concr pt:partb pc:hum:** *горло, рука, лицо, нога*; КАРАНДАШ + **r:concr pt:partb pc:hum:** *рука*; **r:concr pt:partb pc:hum:** *рука* + ЛОПАТА; ТОПОР + **r:concr pt:partb pc:hum:** *рука, лицо, нога*; ЩЕТКА + **r:concr t:hair pt:aggr sc:hair top:rope:** *волос* и т. д.

– части инструментов: **r:concr t:thing pt:residpart pc:X der:v:** *огрызок* + КАРАНДАШ и т. д.

Для большинства существительных — названий инструментов характерны синтаксические группы с фиксированными лексическими элементами, например, **r:rel t:manner:** *как* + БРИТВА, ВЕНИК, ЛОПАТА, ВЕСЛО и т. п. (сравнительный оборот). В данном случае имеется двучленное свободное сочетание структуры *CONJ* + *S*. В выборках зарегистрированы трех- и четырехчленные фраземы: **t:move ca:caus d:root | d:root:** *гнать* + **r:rel ev:neg:** *поганый* + МЕТЛА (*гнать поганой метлой*); **r:concr t:org:** *фирма* + ВЕНИК + *не* + **t:impact:creat d:root:** *вязать* (*фирма веников не вяжет*).

В завершение осуществляется процедура верификации результатов: полученные списки конструкций сопоставляются со списками коллокаций, формируемыми на основе сервиса поиска биграмм С. А. Шарова (**http://corpus.leeds.ac.uk/ruscorpora.html**), с данными словарей МАС и БАС. Обнаружено, что отдельные варианты лексико-семантического наполнения конструкций характеризуются высокой устойчивостью. Среди контекстных маркеров значений целевых слов в составе конструкций присутствуют компоненты коллокаций с высоким показателем Log-Likelihood (LL). Отдельные коллокации нашли отражение в словарях МАС и БАС как устойчивые сочетания (приводятся с пометой ◊ в БАС) и фразеологизмы (приводятся с пометой ◊ МАС и с пометой ~ в БАС) (см. табл. 4).

Таблица 4. Устойчивые сочетания в составе конструкций со словом ***карандаш***

КАРАНДАШ ‘ручное орудие для копания, сгребания с рукояткой и широким плоским отточенным концом’ (r:concr t:tool:instr)		
Конструкции	**Сервис поиска биграмм С. А. Шарова**	**Словари**
r:qual t:physq:color der:s dt:color dt:abstr: *цветной* + КАРАНДАШ		***Цветные*** *карандаши.*
r:concr t:thing pt:residpart pc:X der:v: *огрызок* + КАРАНДАШ		***Огрызок*** *карандаша.*
r:rel der:s: **химический**+КАРАНДАШ	***химический*** LL = 60,34	***Химический*** *карандаш.*
КАРАНДАШ + **d:root:** *писать*		***Писать*** *карандашом.*
КАРАНДАШ + **r:concr pt:partb pc:hum:** *рука*	***рука*** LL = 36,63	

5. Заключение

Исследование показывает, что задача автоматического выделения конструкций, понимаемых как сочетания целевых слов и контекстных маркеров их значений, выражаемых тегами в многоярусной разметке НКРЯ, имеет практическое решение:

(1) разработан инструмент автоматического выделения конструкций, продуктивность которого подтверждена в сериях экспериментов с различными русскими существительными;

(2) разработана и апробирована методика анализа данных о конструкциях, предложены принципы их систематизации по морфологическим моделям, лексико-семантическому наполнению и ряду других критериев;

(3) проведены пилотные эксперименты по выделению конструкций для русских существительных, получены рабочие списки конструкций для

инструментальных существительных, существительных со значением речи и некоторых других слов;

(4) проведена верификация экспериментальных данных: полученные в ходе исследования списки конструкций сравнивались со списками коллокаций для целевых слов, а также с перечнями устойчивых сочетаний и фразеологизмов из словарей русского языка.

Тем самым, лингвистические данные и программные решения, полученные в ходе работы над настоящим проектом, открывают возможность создания каталога русских конструкций, соотносимых с определенными значениями целевых слов.

References

1. *Apresjan Ju. D.* (ed.) (2003) Novyj objasnitel'nyj slovar' sinonimov russkogo jazyka [New Explanatory Dictionary of Russian Synonyms]. Moscow.
2. *Apresjan Ju. D.* (ed.) (2010) Prospekt aktivnogo slovarja russkogo jazyka. [The Prospect of Active Dictionary of the Russian Language]. Moscow.
3. *Bolshakova E. I., Baeva N. V., Bordachenkova E. A., Vasil'eva N.E., Morozov S. S.* Leksiko-sintaksicheskie shablony v zadachah avtomaticheskoj obrabotki teksta [Lexico-Syntactic Patterns for Automatic Text Processing]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog–2007» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2007»]. Moscow, 2007.
4. *Borisova E. G.* (1995) Kollokatsii. Chto eto takoje i kak ih izuchat? [Collocations. What is it and how should we study them?]. Moscow.
5. *Dąbrowska E.* (2004) Language, Mind and Brain: Some Psychological and Neurological Constraints on Theories of Grammar. Edinburgh University Press, Edinburgh and Georgetown University Press, Georgetown.
6. *Fillmore Ch. J.* (1988) The Mechanisms of Construction Grammar. Proceedings of the Berkeley Linguistic Society. Vol. 14.
7. *Gel'buh A. F., Sidorov G. O., èrnandes-Rubio è., Chubukova M. V.* Slovari sochetaemosti slov: kakoj metod sostavlenija luche? [Dictionaries of Word Co-occurrence: Which Way of Building is Better?]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog–2004» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2004»]. Moscow, 2004.
8. *Goldberg A. E.* (1995) Constructions. A Construction Grammar Approach to Argument Structure. Chicago, IL/London: University of Chicago Press.
9. *Goldberg A. E.* (2006) Constructions at Work: the Nature of Generalization in Language. Oxford: Oxford University Press.
10. *Iordanskaja L. N., Melchuk I. A.* (2007) Smysl i sochetajemost v slovare [Sense and Co-occurrence in a Dictionary]. Moscow.
11. *Jagunova E. V., Pivovarova L. M.* (2011) Ot kollokatsij k konstruktsijam [From Collocations to Constructions], Russkij jazyk: konstruktsionnye

i leksiko-semanticheskie podhody [Russian: Constructional and Lexical-Semantic Approaches]. Saint-Petersburg.
12. *Kustova G. I.* Konstruktsii s abstraktnymi sushchestvitel'nymi i ih otrazhenie v èlektronnom slovare [Constructions with Abstract Nouns in an Electronic Database]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog — 2011» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2011»]. Moscow, 2011.
13. *Kuznetsova Ju. L.* (2007) Grammatika konstruktsij. Obzor [Construction Grammar. State of the Art], Nauchno-tehnicheskaja informatsija [Science and Technical Information], Serija 2, № 4.
14. *Lyashevskaya O. N., Kuznetsova Ju. L.* Russkij FrejmNet: k zadache sozdanija korpusnogo slovarja konstruktsij [Russian FrameNet: Towards a Corpus-Based Dictionary of Constructions]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog — 2009» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2009»]. Moscow, 2009.
15. *Lyashevskaya O., Mitrofanova O., Grachkova M., Romanov S., Shimorina A., Shurygina A.* Automatic Word Sense Disambiguation and Construction Identification Based on Corpus Multilevel Annotation. Text, Speech and Dialogue. Proceedings of the 14th International Conference TSD 2011, Pilsen, Czech Republic, September 1–5, 2011. Springer-Verlag, 2011.
16. *Manning C., Schütze H.* (2002) Collocations, in Foundations of Statistical NLP.
17. *Mel'chuk I. A.* (1999/1974) Opyt teorii lingvisticheskih modelej «Smysl <=> Tekst» [On the Theory of Linguistic Models «Sense <=> Text»]. Moscow.
18. *Mel'chuk, Igor A.* (2007) Lexical Functions, H. Burger, D. Dobrovol'skij, P. Kühn & N. Norrick (eds.), Phraseology. An International Handbook of Contemporary Research, 119–213. Berlin/New York: W. de Gruyter.
19. *Mitrofanova O., Lyashevskaya O.* Disambiguation of Taxonomy Markers in Context: Russian Nouns. 17th Nordic Conference of Computational Linguistics NODALIDA — 2009, Odense, Denmark, May 14–16, 2009.
20. *Mitrofanova O. A., Belik V. V., Kadina V. V.* Korpusnoe issledovanie sochetaemostnyh predpochtenij chastotnyh leksem russkogo jazyka [Corpus Analysis of Selectional Preferences of Frequent Words in Russian]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog — 2008» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2008»]. Moscow, 2008.
21. *Mitrofanova O. A., Lyashevskaja O. N., Panicheva P. V.* èksperimenty po statisticheskomu razresheniju leksiko-semanticheskoj neodnoznachnosti russkih imen sushchestvitel'nyh v korpuse [Experiments on Statistical Word Sense Disambiguation of Russian Nouns in the Corpus]. Trudy mezhdunarodnoj konferentsii «Korpusnaja lingvistika–2008» [Proceedings of the International conference «Corpus linguistics — 2008»]. Saint-Petersburg, 2008.
22. *Oskol'skaya S. A, Say S. S.* (2010) Kruglyj stol "Russkij jazyk: konstruktsionnyje i leksiko-semanticheskie podhody" [Workshop on Constructional and

Lexico-Grammatical Approaches to Russian, 2009], Voprosy jazykoznanija [Linguistic Inquiries], Vol. 1.
23. *Ovsjannikova M. A., Oskol'skaya S. A, Say S. S.* (2011) Russkij jazyk: konstruktsionnyje i leksiko-semanticheskie podhody [Constructional and Lexico-Grammatical Approaches to Russian], Voprosy jazykoznanija [Linguistic Inquiries], Vol. 5.
24. *Proceedings* of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics. Los Angeles, CA, 2010.
25. *Rakhilina E. V.* (ed.). (2010) Lingvistika konstruktsij [Construction Linguistics]. Moscow.
26. *Rakhilina E. V., Kobritsov B. P., Kustova G. I., Lyashevskaja O. N., Shemanaeva O. Ju.* Mnogoznachnost' kak prikladnaja problema: leksiko-semanticheskaja razmetka v Natsional'nom korpuse russkogo jazyka [Semantic Ambiguity as an Application-Oriented Problem: Word Class Tagging in the RNC]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog — 2006» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2006»]. Moscow, 2006.
27. *Sahlgren M., Knutsson O.* (2009) Workshop on Extracting and Using Constructions in NLP. NODALIDA'09. SICS Technical Report T2009:10.
28. *Shimorina A., Grachkova M.* Identification of Context Markers for Russian Nouns. 18th Nordic Conference of Computational Linguistics NODALIDA — 2011, Riga, Latvia, May 11–13, 2011.
29. *TKS* — Mel'chuk I. A., Zholkovskij A. K. i dr. (1984) Tolkovo-kombinatornyj slovar' sovremennogo russkogo jazyka. Opyty semantiko-sintaksicheskogo opisanija russkoj leksiki [Explanatory Combinatorial Dictionary og Modern Russian. Semantic-Syntactic Studies of Russian Vocabulary]. Vienna: Wiener Slavistischer Almanach.
30. *Tomasello M.* (2003) Constructing a Language: A Usage-Based Approach to Child Language Acquisition. Cambridge, MA: Harvard University Press.
31. *Wible D., Tsao N.-L.* StringNet as a Computational Resource for Discovering and Investigating Linguistic Constructions. Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics, Los Angeles, CA, 2010.
32. *Zakharov V. P., Khokhlova M. V.* Analiz effektivnosti statisticheskih metodov vyjavlenija kollokatsij v tekstah na russkom jazyke [The Study of Effectiveness of Statistical Measures for Collocation Extraction on Russian Texts]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog–2010» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2010»]. Moscow, 2010.

ВЛИЯНИЕ РЕЧЕВЫХ ХАРАКТЕРИСТИК ДИКТОРОВ НА ТОЧНОСТЬ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ

Людовик Т. В. (tetyana.lyudovyk@gmail.com)

Международный научно-учебный центр информационных технологий и систем, НАН Украины и МОН Украины, Киев, Украина

Исследуется влияние речевых сбоев, темпа речи и других характеристик на точность распознавания украинской подготовленной и спонтанной речи. Обсуждается возможность предсказания точности распознавания речи дикторов исходя из их опыта публичных выступлений. Предлагается область применения системы распознавания речи.

Ключевые слова: автоматическое распознавание речи, украинский язык, спонтанная речь, речевые характеристики

1. Введение

Дикторонезависимое распознавание слитной речи с большим словарем (speaker independent large vocabulary continuous speech recognition) позволяет решать широкий спектр задач, таких как транскрибирование (субтитрирование, стенографирование) аудиозаписей телепередач [1], заседаний парламента [2], судебных заседаний [3] и т. п. Сложность этих задач обусловлена тем, что распознаваемая речь может быть как подготовленной (заранее обдуманное выступление), так и неподготовленной (чтение «с листа», спонтанная речь, например, подсудимых, свидетелей и т. д.). В некоторых случаях распознаваемая речь может быть эмоционально окрашенной (выступления некоторых политиков, а также речь адвокатов и других участников судебных заседаний). Иногда фрагменты подготовленной речи чередуются со спонтанными диалогами и полилогами.

Для решения подобных задач системы распознавания речи должны быть устойчивы не только к техническим факторам (шумовые помехи, свойства каналов передачи акустических сигналов), но и к речевым сбоям, вызываемым ситуацией и индивидуальными речевыми особенностями дикторов [4–7]. К последним относятся оговорки, запинки, паузы хезитации, изменение темпа речи, а также нелитературная лексика (в первую очередь имеется

в виду суржик и сленг). Кроме этого, в речи многих дикторов присутствуют такие паралингвистические явления как вдох/выдох, кашель, смех, причмокивание и т. п.

Подготовленная речь (например, новости, читаемые дикторами телевидения) в настоящее время может распознаваться с пословной точностью 85–90 % [8]. Для этого требуется предварительное обучение системы распознавания на материале больших речевых корпусов (от нескольких сотен до нескольких тысяч часов записанной речи). Распознавание спонтанной речи — значительно более сложная задача, требующая не только соответствующего материала для обучения, но и учета особенностей такой речи при разработке акустико-фонетических и лингвистических моделей систем распознавания.

Анализ проблем, связанных с особенностями спонтанной речи [9], показал, что одной из основных причин ошибок распознавания являются речевые сбои, к которым относятся оговорки, паузы хезитации, а также слишком акцентированное и/или растянутое произнесение отдельных слов. При этом наличие в речи оговорки привело к ошибке распознававния в 100 % случаев, паузы хезитации — в 92 %, растянутого произнесения — в 81 %.

Разработчики из LIMSI (Франция) провели эксперимент со своей системой распознавания речи [1], целью которого было выявить степень влияния речевых сбоев на точность распознавания. Материалом для эксперимента послужили записанные спонтанные телеинтервью на французском языке. Точность распознавания составила 85.5 %. Речевые сбои явились причиной ошибки в 12.5 % случаев. Основной причиной ошибок (25.1 %) явилось редуцированное произнесение слов, свойственное спонтанной речи.

Темп речи, несомненно влияющий на точность распознавания, а также косвенно связанный с речевыми сбоями, исследовался в [9, 10]. В целом, темп спонтанной речи варьирует больше, чем темп подготовленной речи.

Связь речевых сбоев с возрастом и полом дикторов исследовалось на материале английского [4] и венгерского [11] и языков. Подтвердилось предположение о том, что в детском и пожилом возрасте плавность речи нарушается чаще, чем в среднем возрасте. Что касается различия полов, то вопрос о том, кто более склонен к речевым сбоям — мужчины или женщины — остается открытым.

Процентное содержание речевых сбоев в речи может сильно колебаться: от 1.33 % ([12]) до 5.97 % ([4]).

Интересные результаты были получены при исследовании речи, обращенной к компьютеру [12]. Эксперимент показал, что в речи людей меньше сбоев наблюдается при общении с компьютером (0.78 % — 1.87 %), чем при общении с другими людьми (5.50 % — 8.83 %). Возможно, это связано с более медленным темпом, а также большей мотивацией продумать (спланировать) реплику от начала до конца. По крайней мере, результаты данного исследования говорят о том, распознавание речи в человеко-компьютерных системах должно быть не сложнее, чем распознавание спонтанной речи, адресованной людям.

2. Цель исследования

Целью работы является исследование влияния ситуативных и индивидуальных факторов на точность распознавания речи. Под ситуативными факторами в данной работе понимаются условия протекания речевой коммуникации (теледебаты, интервью, чтение теленовостей, судебный процесс) и роли коммуникантов. К индивидуальным факторам отнесены степень подготовленности к публичным выступлениям, склонность к речевым сбоям, степень употребления нелитературной лексики.

Полученные результаты могут быть использованы для предсказания точности распознавания речи при появлении новых дикторов и/или новых приложений системы.

Полученная информация позволит определить сферы применения системы распознавания речи исходя из результатов, достигнутых на текущий момент.

Предполагается определить степень влияния различных индивидуальных характеристик дикторов на точность распознавания их речи. В частности, внимание уделяется речевым сбоям, темпу произношения, паралингвистическим явлениям и употреблению нелитературной лексики (суржика).

Конечной целью работы является повышение точности распознавания как подготовленной, так и спонтанной речи.

3. Речевой материал

Исследованный речевой материал, содержащий 16 699 реализаций словоформ, был взят из Акустического корпуса украинской эфирной речи (АКУЕМ) [13] (86 %), непосредственно из телеэфира (11 %) и записан в студии (3 %). Тип сигнала — моно, частота дискретизации 22,05 кГц с 16-битным разрешением, формат Windows PCM wav.

Аудиозаписи корпуса АКУЕМ, разработанного фирмой «Специальные регистрирующие системы» (Киев, Украина), имеют аннотацию, включающую орфографический текст и сведения о речевых сбоях (оговорки, паузы хезитации), паралингвистеских явлениях (вдох/выдох, кашель, смех, плач), а также о значительно редуцированном или растянутом произнесении. Кроме этого, в корпусе АКУЕМ отмечены такие явления как суржик, жаргон, диалектизмы.

По содержанию исследованный материал представляет собой новостные телепередачи (15 %), ток-шоу (29 %), телепередачи судебной тематики (25 %), а также речь, прозвучавшую в реальных судебных заседаниях (31 %). Таким образом, представлена подготовленная речь (чтение новостей), слабо подготовленная речь (выступления политиков, обвинительные речи прокуроров), неподготовленное чтение (зачитывание судебных постановлений) и спонтанная речь (в основном, опросы свидетелей и судебные прения).

Исследовалась речь дикторов телевидения, политиков, журналистов, юристов и «рядовых» участников судебных процессов. Женщин и мужчин было

поровну. В Таблице 1 приведены сведения о дикторах. Заглавная буква кода соответствует роду занятий диктора или его роли в речевом общении (Ю — юрист, П — политик, Д — диктор телевидения, Ж — журналист, С — свидетель (возможно также подсудимый или потерпевший) в телепередаче судебной тематики. Следует подчеркнуть, что в роли судей, прокуроров и адвокатов выступают реальные юристы. Строчная буква кода обозначает пол диктора.

Особый интерес представляет речь юриста Ю_ж_1, поскольку исследовались записи реальных судебных заседаний с ее участием. В одних случаях судья опрашивала подсудимых и свидетелей (спонтанная речь), в других — оглашала протоколы заседаний (чтение).

Таблица 1. Распределение дикторов в соответствии с их родом деятельности и полом

Коды дикторов	Д_ж	Д_м	П_ж	П_м	Ж_ж	Ж_м	Ю_ж	Ю_м	С_ж	С_м	Всего
Количество дикторов	2	1	1	5	2	1	2	1	2	1	18

4. Система распознавания украинской речи

Система распознавания украинской речи [2, 3] разработана на основе инструментария HTK [14].

В качестве акустических моделей (АМ) используются скрытые Марковские модели 70 фонем, обученные на материале корпуса АКУЕМ (82 часа аннотированной речи, 1150 дикторов). Для обучения был отобран речевой материал, не содержащий таких помех как одновременное звучание нескольких голосов, звук аплодисментов и другие не зависящие от диктора шумы. Помимо акустических моделей 53 фонем украинского языка используются особые модели для паузы, звучащих пауз хезитации («э-э-э», «а-а-а», «м-м»), вдохов/выдохов, чмоканья, кашля, смеха и плача.

Лингвистическая модель (ЛМ) представляет собой биграммную модель языка, заданную вероятностями пар словоформ. ЛМ обучена на текстах корпуса АКУЕМ, текстах из Интернета и нескольких искусственно созданных текстах, в которых представлены все комбинации «день+месяц» и часть комбинаций «день+месяц+год» (например, в переводе с украинского: «первое января», «первого января», «второго февраля две тысячи двенадцатого года»).

Словарь распознавания содержит 115 900 словоформ. Часто употребляемые в речи словоформы, в том числе числительные, имеют от 1 до 10 фонемных транскрипций, остальные — от 1 до 3. В среднем на одну словоформу в словаре распознавания приходится 1.55 фонемных транскрипций, отражающих как типичное, так и спонтанное произнесение. Основная часть транскрипций порождена автоматически. Так, 3924 транскрипции числительных были созданы автоматически, а 34 добавлены вручную (ср. русское «тысячи» [т Ы ш’ и]).

5. Эксперименты

Было проведено два эксперимента, один из которых состоял в распознавании речи, а второй основывался на корпусном исследовании.

5.1. Распознавание речи

С помощью системы распознавания украинской речи были получены данные о точности распознавания речи 18 дикторов. Пословная точность распознавания вычислялась с помощью инструментария HTK путем сравнения «правильного» (эталонного) текста (reference) с ответом распознавания в виде последовательности словоформ. Учитывались замены, вставки и пропуски словоформ. Не учитывались вдохи/выдохи и хезитации.

5.2. Корпусное исследование

На материале корпуса АКУЕМ были исследованы речевые характеристики, зависящие от диктора и ситуации речевого общения — речевые сбои, темп, редуцированное и растянутое произнесение отдельных слов, а также присутствие в речи нелитературной лексики. Речевой материал, не входящий в корпус, был предварительно аннотирован в соответствии со стандартами корпуса.

Темп речи измерялся количеством произнесенных в секунду слогов. Для каждого диктора экспертом были выбраны характерные участки речи без внутренних пауз. Подсчет слогов производился по гласным звукам (являющимся слогообразующими) соответствующих фонемных транскрипций, полученных автоматически, без учета особенностей произношения дикторов. Полученные значения темпа речи рассматриваются как приблизительные, поскольку и гласные в речевом потоке могут выпадать, и темп в целом может варьировать в зависимости от таких факторов как усталость, переход от чтения к спонтанной речи и т. д.

Речевые сбои были объединены в две категории: хезитации («а-а-а», «м-м», ...) и оговорки (фальстарты и произнесение не того слова, которое планировалось).

Из нелитературной лексики в исследованном материале присутствовал только суржик.

Количественной оценкой речевых сбоев, суржика, а также вдохов/выдохов, редуцированного и растянутого произнесения отдельных слов являлось их процентное отношение к общему количеству слов исследуемого речевого материала диктора.

Полученные результаты были сопоставлены с результатами распознавания и представлены в Таблице 2. Речь диктора Ю_ж_1 представлена как Ю_ж_1(ч) и Ю_ж_1(с) — соответственно как чтение и спонтанная речь.

Таблица 2. Результаты распознавания речи и речевые характеристики дикторов

Дикторы	Количество слов	Точность распознавания	Темп речи	Шумное дыхание	Хезитация	Оговорки	Редукция	Растягивание	Суржик
Д_ж_1	968	89.77	7.7	0	0	0	0	0	0
П_м_1	736	88.32	5.7	0,8	1,5	0	0	0	0
Д_м_1	875	87.09	6.5	2,4	0,6	0,1	0	0	0
П_ж_1	276	84.06	7.3	0,7	0	0,7	0	0	0
Д_ж_2	459	83.66	6.0	0	0	1,5	0	0	0
П_м_3	491	81.06	6.2	0	0,6	0,6	0,2	1,6	1,2
С_ж_1	508	78.94	6.0	1,0	2,6	1,6	0,6	2,4	0
П_м_2	1258	77.90	5.6	0,6	1,3	0,5	0,7	3,1	0
Ю_ж_1(ч)	1553	76.11	7.4	5,6	0,1	0,6	1,5	0,3	0,4
Ю_ж_2	1976	74.80	4.9	1,0	3,0	1,5	0,8	1,0	0,5
Ж_ж_2	169	72.78	6.8	1,2	3,0	1,2	0	0,6	0
Ж_ж_1	255	71.76	6.6	3,9	0	0	1,6	0,4	0
Ю_м_1	1102	69.96	6.2	4,9	0,6	0,6	0,3	1,9	0,5
Ж_м_1	548	69.71	6.3	2,7	3,1	0	3,1	1,3	0
Ю_ж_1(с)	3725	64.35	5.3	5,3	2,8	0,6	0,1	0	2,5
П_м_4	315	62.86	7.3	1,0	0,3	0,3	3,2	1,3	0,3
П_м_5	958	57.83	6.8	0	2,0	0,3	2,6	1,0	0,3
С_м_1	175	47.43	4.7	1,1	5,7	1,7	0	12,6	4,6
С_ж_2	352	46.02	6.7	0	2,0	1,7	1,7	1,4	0,3

6. Обсуждение результатов

Эксперименты показали, что наименьшее влияние на точность распознавания оказывают шумные вдохи/выдохи и хезитации. Это объясняется тем, что данные явления успешно моделируются и распознаются системой [3]. Влияние темпа речи на точность распознавания неоднозначно. Быстрый темп речи дикторов телевидения, обладающих навыками четкой артикуляции, не ухудшает результаты, в то время как более эмоциональная быстрая речь политиков распознается хуже. На материале диктора Ю_ж_1 видно, что быстрая подготовленная речь (чтение) распознается на 12 % лучше, чем медленная спонтанная.

В целом исследуемые речевые характеристики можно разбить на три группы: темп, характеристики, влияющие на точность распознавания (оговорки, редукция, растягивание, суржик) и не влияющие (вдохи/выдохи и хезитации). Эти группы характеристик предсталены на диаграмме (Рис. 1). Горизонтальная ось соответствует списку дикторов, упорядоченному по убыванию точности распознавания речи. По вертикали представлены значения точности (%), а также масштабированные с коэффициентом 10 значения темпа речи

и с коэффициентом 5 суммированные значения влияющих и не влияющих на точность распознавания речевых характеристтик.

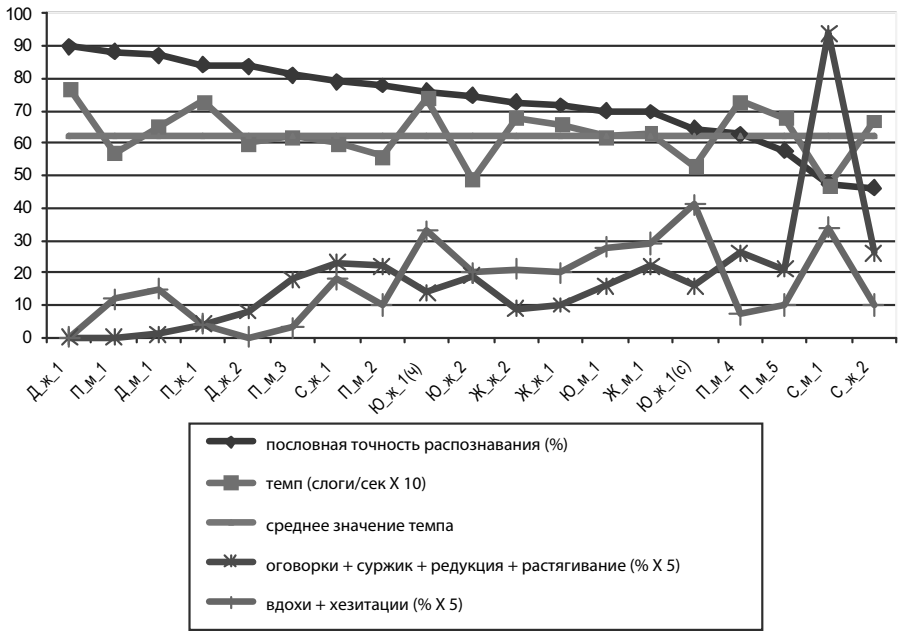


Рис. 1. Точность распознавания речи различных дикторов и их речевые характеристики

Особенно явно различия между группами характеристик заметны на результатах дикторов, чья речь распознана с точностью более 85% (первые три диктора) и с точностью менее 50% (два последних диктора). Речь «отличников», среди которых два диктора телевидения (женщина и мужчина), практически лишена суржика и оговорок. Речь «отстающего» диктора С_м_1 отличается медленным темпом и растянутым произнесением каждого восьмого слова.

Если рассматривать влияние речевых характеристик на точность распознавания речи, то основными причинами ошибок являются: 1) оговорки; 2) использование диктором нелитературной лексики; 3) редуцированное произнесение слов; 4) растянутое произнесение слов. Первые две причины приводят к тому, что «не срабатывает» лингвистическая модель системы распознавания, основанная на статистических закономерностях употребления слов. Третья и четвертая причины вызывают проблемы, связанные с акустической моделью системы распознавания речи.

Результаты распознавания украинской речи подтверждают вывод, сделанный нами ранее на материале распознавания русской речи: хорошо распознается речь дикторов, имеющих большой опыт публичных выступлений. Закономерно, что лучшие результаты по точности распознавания у дикторов

телевидения, а также политиков, имеющих опыт работы пресс-секретарем, главой администрации, спикером. Причем, если в исследуемом материале речь профессиональных дикторов — это чтение текста, то речь политиков — речь не «по бумажке» (в основном ответы на вопросы).

Полученные результаты позволяют приблизительно предсказать точность распознавания речи диктора исходя из его профессии: профессиональные дикторы радио и телевидения — 80–90 %, политики-«спикеры» — 75–90 %, журналисты, телеведущие и юристы — 60–80 %, остальные — 45–80 %.

Важна также роль диктора в речевой ситуации. Предварительный анализ речи расширенного состава участников судебных заседаний свидетельствует о том, что в целом по точности распознавания речи прокуроры уступают судьям, но превосходят адвокатов. Это может быть объяснено, в частности, тем, что речь судей наиболее нейтральна, а речь адвокатов наиболее эмоциональна, и, следовательно, менее контролируема (в большей степени подвержена речевым сбоям), однако одновременно и более акцентирована на отдельных участках (растягивание слов, послоговое произнесение).

Немаловажным является мотивированность речи, желание диктора, чтобы его слушали и понимали. В этой связи можно упомянуть получающие все большее распространение речевые технологии голосового поиска, перевода, ввода текстов для электронной почты и т. п. Если есть заинтересованность в адекватной реакции персонального голосового помощника, с которым происходит устное общение, то необходимо произносить голосовые запросы и команды четко (но без гиперартикуляции) и в стабильном темпе. Можно предположить, что голосовое общение с техническими устройствами будет способствовать совершенствованию речевых навыков (своего рода «персональный логопед»).

Исследования, проведенные в данной работе, будут в дальнейшем продолжены в направлении автоматизации выявления значений речевых характеристик и совершенствования подходов их моделирования в системе распознавания речи.

7. Выводы

Точность распознавания речи зависит от того, насколько успешно моделируются в системе распознавания речи характеристики, обусловленные ситуативными и индивидуальными факторами.

Основными речевыми характеристиками, влияющими на точность распознавания речи, являются: 1) оговорки; 2) использование диктором нелитературной лексики; 3) редуцированное произнесение слов; 4) растянутое произнесение слов.

Фактором, положительно сказывающимся на точности распознавания речи, являются навыки устных выступлений. Полученные результаты позволяют приблизительно предсказать точность распознавания речи диктора исходя из его профессии.

Точность распознавания украинской речи, достигнутая к настоящему времени, позволяет использовать систему распознавания речи для автоматического субтитрирования новостных телепередач.

Литература

1. *Adda-Decker M., Habert B., Barras C., Adda G., Boula de Mareuil P., Paroubek P.* A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models. Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop, Göteborg, Sweden, 2003, pp. 67–70.
2. *Пилипенко В. В., Робейко В. В.* Автоматизированный стенограф украинской речи // Искусственный интеллект. Донецк: 2008. № 4. С. 768–775. Pilipenko V. V., Robeiko V. V. (2008), Automated stenographer of Ukrainian speech [Avtomaticheskii stenograf ukrainskoi rechi], Iskusstvennyi intellekt [Artificial Intelligence], no. 4, pp. 768–775.
3. *Людовик Т. В. Пилипенко В. В., Робейко В. В.* Автоматическое распознавание спонтанной украинской речи (на материале акустического корпуса украинской эфирной речи) // Компьютерная лингвистика и интеллектуальные технологиии. По материалам ежегодной Международной конференции «Диалог 2011». Вып. 10. М.: РГГУ, 2011. С. 478–488. Lyudovyk T. V., Pilipenko V. V., Robeiko V. V. Automated stenographer of Ukrainian speech (on material of acoustic corpus of Ukrainian speech) [Avtomaticheskoe raspoznavanie spontannoi ukrainskoi rechi (na materiale akusticheskogo korpusa ukrainskoi efirnoi rechi], Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2011" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011"]. Bekasovo, 2011, pp. 478–488.
4. *Bortfeld H., Leon S. D., Bloom J. E., Schober M. F., Brennan S. E.* (2001), Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender, Language and Speech, Vol. 44 (2), pp. 123–147.
5. *Meyer B. T., Brand T., Kollmeier B.* (2011), Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes, Journal of the Acoustical Society of America, Vol. 129, Issue 1, pp. 388–403.
6. *Choularton S.* (2007), "Early Stage Detection of Speech Recognition Errors", available at: http://www.xesoftware.com.au/ThesisAsPassed.pdf
7. *Greenberg, S., Chang, S.* Linguistic dissection of switchboard-corpus automatic speech recognition systems. Proceedings of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium. Paris, 2000, pp. 195–202.
8. *Le* Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI) (2010), Rapport d'activité 2010, available at: http://www.limsi.fr/Rapports/LIMSI_Rapport2010.pdf.

9. *Butzberger J. W., Murveit H., Shriberg E. E., Price P. J.* Spontaneous speech effects in large vocabulary speech recognition applications. Proceedings of the DARPA Speech and Natural Language Workshop. San Mateo, CA, 1992, pp. 339–343.
10. *Trouvain J., Koreman J., Erriquez A., Braun B.* Articulation rate measures and their relation to phone classification in spontaneous and read German speech. Proceedings of ITRW on Adaptation Methods for Speech Recognition Adaptation-2001. Sophia Antipolis, France, 2001, pp. 155–158.
11. *Menyhart K.* Age-dependent types and frequency of disfluencies. Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop. Goteborg University, Sweden, 2003, pp. 45–48.
12. *Oviatt S.* (1995). Predicting spoken disfluencies during human-computer interaction. Computer Speech and Language, no. 9, pp. 19–35.
13. *Pylypenko V., Robeiko V., Sazhok M., Vasylieva N., Radoutsky O.* Ukrainian Broadcast Speech Corpus Development. Proceedings of SPECOM 2011, 14-th International Conference "Speech and Computer". Kazan, Russia, 2011, pp. 435–440.
14. *Young S.* et al. (2009), "The HTK Book (for HTK Version 3.4)", available at: http://htk.eng.cam.ac.uk/.

ORAL SPEECH AND INSTANT MESSAGING: THE IMPACT OF THE COMMUNICATION CHANNEL ON DIALOGUE STRUCTURE

Liusina V. S. (krotcaja@gmail.com)

Institute of Linguistics of the Russian Academy of Sciences, Moscow, Russian Federation

Correlation between oral and virtual dialogues is analyzed. It is known that dialogues consist of minimal dialogic units (MDU). Contrary to a widespread opinion that an MDU cannot include another MDU, there are evidences that this is possible when one of the dialogues is not intentional. The structure of 60 Russian Internet dialogues is analyzed. If these dialogues satisfy the "close-to-oral-conversation" conditions (i.e. the interlocutors are not involved in any other mental activities while talking), a mutation of dialogue structure occurs. There are five steps of mutation that lead from an overlooking of a remark to a communicative failure. In such "mutated" dialogues, two interlaced "intentional" MDU are present. However, the communicative failure occurs very rarely due to multitasking. Normally, people are involved in several activities while talking via Internet, which provides necessary pauses in Internet conversation. It can be concluded, therefore, that Internet conversation is a wholly new type of communication.

Key words: instant messaging, dialogue structure, minimal dialogical unit, communicative failure

1. Introduction

In this article, the problem of everyday dialogue via instant message clients is investigated. This type of dialogue is often referred to as "chat", however, I will not use this term due to its homonymous nature. First, the relevant features of oral dialogues will be discussed; then, a comparison of oral and virtual dialogues will be made.

2. Oral dialogue

Not so long ago, everyday dialogue was possible only via oral communication (tête-a-tête or by phone), but in the last two decades it has increasingly occupied virtual space. Nevertheless, we cannot affirm that this is still the same dialogue, although coded in a different way. As we will see further, the communication channel influences the dialogue structure enormously, converting instant message conversations into a whole new type of communication.

Most researchers of oral dialogue focus on its outer characteristics, pointing out that dialogue represents a collaboration of communicants [Sacks et al. 1974]; others, on the other hand, investigate remarks in dialogue as a realization of various speech acts [Traum 1999]. In [Baranov, Kreidlin 1992], an attempt was made to reveal inner mechanisms of dialogue by showing the relations between remarks that constitute it without any linkage to concrete speech acts.

The authors construct a definition of a minimal dialogic unit (henceforth MDU) on the basis of the notion of illocutionary force. According to them, MDU is a text that

- starts with an absolutely free remark (that is not forced by any other remark);
- ends with an absolutely bounded remark (that doesn't force any other remark);
- has only one subject matter;
- satisfies all the conditions of illocutionary force.

Moreover, it is said that an MDU cannot include another MDU.

In oral speech, the last criterion is not obligatory. For example, in [Zemskaia et al. 1983] the coexistence of several themes is regarded as a particular characteristic of oral speech. As an example, the following dialogue is given:

(1) *A. Podlej nemnogo// I on mne govorit// B. (l'et moloko v kastrjul'ku) A. Hvatit-hvatit// Ja skoro opjat' priedu. B. A ne malo budet? A. Net-net// B. Kogda zhe/ skoro? V budushchem godu? A. Da net! V ètom eshche//*
'A: Pour a little more// And he is saying to me// B. (is pouring milk in a pot) A. That's enough// I will come again soon. B. Won't it be not enough? A. No, no// B. So when/ will it be soon? Next year? A. Oh, no! This year//'

In this case, two MDU can be distinguished. The free remark of A is forcing B to act; this action can be regarded as a forcing remark, as it produces A's next remark *Hvatit-hvatit*. Let us note that this non-verbal component literally "interrupts" A, forcing her to pause in her remark. Later, B returns to the first subject of the conversation by asking *A ne malo budet?*, and a goes back to the second subject only after satisfying B's illocutionary intentions.

Thus, both dialogues start with absolutely free remarks and end with absolutely bounded remarks; they satisfy the illocutionary force conditions and both are devoted to one and only one particular subject. That is, both are undoubtfully MDU. However, one dialogue is incorporated into the other, which means that the last MDU criterion is not satisfied.

This is quite understandable, as in [Baranov, Kreidlin 1992] "crystallized" fiction dialogues are investigated. It has to be noted that dialogues similar to the above-mentioned occur only if one subject of the conversation is more "intentional", and another is more "pass-through". That is, the communicants have an intention to talk only on one subject, but if they are talking while doing something simultaneously, their actions sometimes need a commentary. In this case, a parallel dialogue "in parenthesis" occurs. In such a dialogue we expect a number of non-verbal remarks, and normally a non-verbal remark plays a forcing role in respect to the verbal ones. Consider the following example:

(2) *A: I tut ja emu govorju// Peredaj, pozhalujsta, sol'.// A pochemu by vam ne pogovorit' s nej samomu?// Net-net, druguju, èto perets. Spasibo.// B: I on chto? A. Obidelsja, konechno.*
'A: And then I tell him // Pass the salt, please. // Maybe you should talk to her yourself? // No, the other one, that's the pepper. Thank you. // B: And what about him? // A: He was offended, of course.'

We have here two interlaced MDU, the principal "intentional" one refers to some events that happened to A in the past, the one "in parentheses" is dedicated to the actual "here-and-now" needs of A. Although in the "salt MDU" only A's remarks are verbalized, we can postulate two zero lexemes that represent non-verbal remarks of B:

(3) *A: I tut ja emu govorju// Peredaj, pozhalujsta, sol'. — B: Ø.// A pochemu by vam ne pogovorit' s nej samomu?// Net-net, druguju, èto perets. — B: Ø. — Spasibo.//*

Let us go back to (1). Intuitively, we can say that A:
a) was expecting a non-verbal answer to the firs remark and therefore introduced a new subject right away;
b) was expecting B to finish pouring milk without any additional remarks; since this did not happened, A was forced to interrupt herself;
c) was not expecting B to ask any additional questions on the "milk topic".

Same thing can be noted about (2): A wasn't intending to extend the "salt topic", but B's mistake forced her to do so.

To sum up, an MDU represents an exchange of remarks, of which every remark is connected to the previous one, the next one or both. Normally, an MDU doesn't include another MDU. If this happens, the incorporated MDU demonstrates the following characteristics:
1) The initiating remark is a request or an order.
2) The second communicant's replies are non-verbalized.
3) There is no intention to extend or develop this dialogue (normally on both sides).
4) If the number of remarks is more than three, this means that one communicant doesn't satisfy the other.

In section 4, we will see if these rules work for virtual communication.

3. Instant messaging

In the last decade, many works on internet communication and its connection to oral communication have been published. One of the most relevant to our study is [Herring 1999], where incoherence of Internet communication is investigated.

Herring points out two problems of instant messaging that are not characteristic for verbal communication: lack of simultaneous feedback and disrupted turn adjacency. This means that in a conversation via oral communication the production of a remark and the perceiving of it are simultaneous, while in instant messaging the addressee starts perceiving the remark only after the speaker has finished it. Because of this, one communicant may introduce another topic and then get an answer to the previous one, which leads to disrupted turn adjacency, that is, to the remoteness between remarks that are connected to each other.

Although it contains a number of valuable observations, [Herring 1999] is now out-of-date. The facilities of Internet communications have grown enormously since 1999; moreover, Herring analyzed the data of group chats which don't play an important role nowadays.

In the next few years the problem of coherence in instant messaging has been extensively discussed in such works as [Baron 2009], [Voida et al. 2002], [Lam&Mavkiewicz 2007], [Berglund] etc. It has been noted that instant messaging creates problems because of its intermediate status between oral and written speech [Baron 2009, Voida et al. 2002]. It has been pointed out that besides the lack of simultaneous feedback and disrupted turn adjacency, the lack of constant attention creates an important problem for conversational coherence. The constant record of a conversation has been acknowledged as an advantage [Herring 1999, Berglund 2010]. Moreover, techniques that are used to maintain conversational coherence have been investigated. The most important one is maybe the dividing of a remark into several parts which Baron [Baron 2009] compares to the "intonation units" of oral speech. Moreover, topicalization and performative verbs are widely used [Lam&Mackiewicz 2007]. In [Berglund 2010] it is shown that, although 144 of 1689 remarks are characterized by disrupted turn adjacency, only 18 of them cause communication problems. The reported techniques are substitution, repetition and conjunction.

The results of these investigations are quite valuable, but the data used may not be totally appropriate. Firstly, the subjects were aware of the fact that their conversations will be investigated, which might have influenced their manner of constructing a dialogue. Secondly, in some cases English was not a native language for the subject, which might have made the conversations too short and primitive. Finally, the subjects were often not so familiar with each other, and didn't have an experience of oral dialogue; therefore, it could be not quite accurate to compare these conversations to oral dialogues.

In the next section, I offer my own investigation of Internet dialogues.

4. Structure of Internet dialogues

For this study, 60 Google Talk dialogues in Russian were analyzed. The data was provided by five communicants. For all of them, Russian is their native language. After my request, the subjects have chosen several dialogues that they could let analyze and publish. The records are ordinary dialogues between friends that

communicate orally on a regular basis. There are 25 short dialogues (less than 50 lines), 15 medium dialogues (between 50 and 150 lines) and 20 long dialogues (more than 150 lines).

If we want our comparison of oral and Internet dialogues to be accurate, we have to exclude all Internet conversations that differ from oral conversations not only in the channel of communication, but also in some other ways. That is, we have to disregard Internet dialogues that were accompanied by reading, writing or talking to other people via Internet. However, this task is too difficult. If we want the dialogues to be spontaneous and natural, we shouldn't tell anything to the subjects before the conversations; it is therefore impossible to create "similar-to-oral-conversation" conditions that we need. When we collect the data post factum, we cannot find out if the subjects were doing anything else while talking, because they don't remember. Moreover, it is very rare that the communicant isn't involved in some other activity while talking. Normally, instant messaging is parallel to some other activities, as some investigators point out [Baron 2010] and as our own experience suggests. However, some dialogue sections in our data are clearly of the "similar-to-oral-conversation" nature. In these sections, there is no noticeable pause between remarks, and frequency is nearly 10 lines in a minute. Moreover, it can be noted (although we don't have a formal instrument to prove it) that the topic of the conversation is too interesting for the subjects to switch to some other activities. It is in such dialogues that the interlacing of two MDU occurs.

In our data, there are 29 dialogues that demonstrate interlacing of two MDU. The analysis has shown that there are five steps that lead to the interlacing and, finally, to the communicative failure. These steps are described below.

1. Overlooking of a remark

In a fast dialogue without pauses, it is common that one communicant misses a remark of the other. As it is a frequent technique to divide remarks into several parts, a reply of the second communicant may be wedged between the parts of a remark of the first communicant. Moreover, a reply of the second communicant may appear after the whole remark of the first one, introducing a topic that has already been introduced. This is illustrated in (4) and (5).

(4) ***A**: […] napisala, chto ni ja ni ty ne v detskom sadu chtoby rasfrenzhivat' i ne v 5 klasse chtoby posylat'*
***B:** da, tak, pozhaluj, luchshe)*
i teper' ty budesh' dal'she stradat' iz-za K?
***A**: koroche govorya kakoj-to tresh polnyj*
'**A**: I wrote that neither me nor you are in kindergarten anymore to be unfriending or in fifth grade to be telling eachother to leave alone
B: yeah, it's probably better this way
So now you're going to continue being miserable about K.?
A: in short, it's all completely crazy'

(5) ***A**: ucheba? zdes' prekrasny frantsuzkie distsipliny i otvratitel'ny ispansistskie*

*audiovizual'nyi **perevod** otlichnyi*
u vas Bea vela?
***B**: a na **perevod** ty hodish'?*
'**A**: School? The French subjects are wonderful here, the hispanist subjects are horrible
the audiovisual **translation** is great
Did you have Bea?
B: Do you take **translation**?'

In (4), what is supposed to be a monologue of A is interrupted by B's remark. A doesn't notice it and keeps talking. Let us note that this situation would never occur in oral speech, where B's remark would be noticed immediately.

(5) is even more striking. A introduces a new topic and adds one more remark (a question *u vas Bea vela?*), but B overlooks both and asks a question on the same topic (*a na perevod ty hodish'?*).

2. Delayed response

After some delay, the first communicant notices the overlooked remark and goes back to it, as illustrated in (6) and (7).

(6) ***A:** i teper' ty budesh' dal'she stradat' iz-za K?*
***B**: koroche govorya kakoj-to tresh polnyj*
neee, chto ty. sovershenno ne budu bol'she stradat'. <...>
***A**: da uzh, tresh tak tresh*
'So now you're going to continue being miserable about K.?
B: in short, it's all completely crazy
Noooo, I'm not going to continue being miserable at all
A: yeah, crazy it is'

(7) ***A**: kak vy vstretili Anyu?*
***B**: horosho*
***A**: mnogo kto byl?*
[...]
***A**:u menya zavtra ocherednaya popytka sdat' drevneshvedskij, poètomu ja reshila ne idti(*
***B**: a pochemu ty ne mogla?*
aaa
a chto, èto tak slozhno?
'**A**: how's you meet Anya?
B: OK
A: were there many people?
A: I'm having another try at passing old Swedish, so I decided not to go
B: why couldn't you?
Ooh
It's that hard?'

If the overlooked remark is close to the main topic of conversation, right turn adjacency is soon restored. However, if a whole new topic is introduced, it may lead to recursion.

3. Recursion

On this stage, the first communicant is replying to the second communicant's remark, while the latter is replying to the overlooked remark of the former. This is illustrated in (8). In line 5, B introduces a new topic, while line 6 continues A's monologue. In line 7, A is going back to the overlooked remark (line 5), while B is replying to the remark introduced in line 6. In line 9, B continues the new topic, and so does A in line 10. The last flash of the "problem topic" is an emoticon introduced by A in line 11. Then the conversation switches to the "cardigan topic".

(8) *1* ***A****: a ja proverjaju zadachku*
2 (ne sejchas)
3 ***B****: ty kruta kak ja ne znaju chto*
4 ***A****: ja byl b kruta, esli by sama sochinila chto-nibud'!*
5 ***B****: ty mne luchshe skaji, s chem ty nosish' svoj seryj kardigan iz frantsii?*
6 ***A****: a poka ja sochinila tol'ko 1 zadachku dlja medvezhonka*
7 vot èto slozhnyi vopros na samom dele
8 ***B****: i tri zadachi dlja treh medvedej*
9 ja kupila sebe pohozhego tipa
10 ***A****: s uzkimi dzhinsami i majkoj, kazhetsja, adekvatnee vsego*
11))
12 ***B****: s majkoi moj ne smotritsja*
13 ***A****: u tebja tozhe seryj?*
14 ***B****: on trehtsvetnyj*
15 2 ottenka serogo i chernyj
16 ochen' krasivyj
17 ***A****: ja tak schitaju, chto ego nado nosit' s plat'ami*
'1 **A**: I'm checking a problem
2 (not at the moment)
3 **B**: you're awesome like I don't know what
4 **A**: I would've been awesome, if I wrote something myself!
5 **B**: tell me, what do you wear with your gray cardigan from France?
6 **A**: I only wrote one problem for the "Bear cub"
7 That's a hard question really
8 **B**: and three problems for three "bears"
9 I bought one sort of like it
10 **A**: tight jeans and a shirt I guess are most appropriate
11))
12 **B**: mine doesn't look good with a shirt
13 **A**: yours is gray too?
14 **B**: it's tricolor
15 2 shades of gray and black

16 Very pretty
17 **A**: I think it should be worn with dresses'

4. Parallelism

When both topics are relevant and important for the communicants, this may result in two parallel dialogues. In (9), remarks of the second topic are less frequent than the first one. However, there are no signs of declension; both topics are developing symmetrically.

(9) *1 **A**: a na dushe-to kak? kakie vetry duyut?*
*2 **B**: tak, vpolne snosno*
3 tol'ko s NG vse kak-to grustno
*4 **A**: a na kursy po ital'janskomu ty eshche ne zapisyvalas'?*
5 a chego s NG?
*6 **B**: n, mozhet, vsplyvet chto-nibud' v poslednij moment*
*7 **A**: v smysle net planov?*
*8 **B**: tak tam zhe vrode v janvare novyj nabor*
9 da
*10 **A**: a my vot dom sobiraemsja snjat' v Podmoskov'je, ajda?*
11 ne na sam NG, pravda
*12 **B**: vy?*
*13 **A**: tipa s 3 po 9, esli poluchitsja konechno*
14 nu tam kompanija narodu
15 a ja nashel dom, tak chto mogu zvat' kogo hochu)
*16 **B**: bylo by ochen' zdorovo, no ne budet nemnogo ne v kassu?*
17 pokazhi!
*18 **A**: am dejstvitel'no, v janvare navernoe, prosto ty ugrozhala pomnju, chto vot-vot zapisyvaesh'sja)*
19 ty tam polovinu lyudej znaesh', ja dumaju
*20 **B**: ne, ja ugrozhala, chto vot-vot nachnu uchit' ital'janskij))*

(10) '1 **A**: what's on your mind? Where's the wind blowing?
2 **B**: it's ok really
3 Except it's kind of sad about New Year's
4 **A**: Have you registered for the Italian course yet?
5 What about new year's?
6 **B**: but maybe something will come up at the last second
7 **A**: you mean you don't have plans?
8 **B**: the new admissions are in January, aren't they?
9 Yeah
10 **A**: we're going to rent a house in the Moscow area, wanna come with us?
11 Not New Year's Eve itself though
12 **B**: who's we?
13 **A**: like from the 3rd to the 9th if it works out of course
14 It's a group of people

15 I found the house, so I can invite whoever I want
16 **B**: it would be great, but won't it be a little awkward?
17 Show me!
18 **A**: oh yeah, it probably is in January. It's just that you threatened me you were about to register
19 You know half the people there I'm sure
20 **B**: No, I threatened that I was about to start studying Italian.'

5. Communicative failure

The situation presented in (9) is a dangerous one. Sooner or later, there comes a moment when one or both of the communicants fail to identify which topic the remark belongs to. This leads, then, to the communicative failure, as in (10). In line 19, A refers to the main topic of the conversation. This is accompanied by a delay from A: in line 20, she replies to the remark presented in the line 18; only then she notices line 19 and by mistake attributes it to the "Italian topic", which is clear from line 21. This is quite understandable, as the remarks before and after presented in line 19 are devoted to this topic. Therefore, a communicative failure occurs. Again, there is a delay: A replies to the remark presented in line 20 and misses line 21. It is in line 23 that he reacts to the question, but his reply doesn't resolve the failure; on the contrary, A doesn't understand B's question, so he replies with another question. The answer of B is also a question, which shows that the communicative failure is still not resolved. Let us note that B doesn't reply to A's question presented in line 22.

20 ***B****: ne, ja ugrozhala, chto vot-vot nachnu uchit' ital'janskij))*
21 ***ljudej na ital'janskom?***
22 ***A****: nu i kak, udalos'?*
23 ***chto ljudej na ital'janskom?***
24 ***B****: chto udalos'? ja chto-to terjaju nit'*
'20 **B**: No, I threatened that I was about to start studying Italian.
21 **People in Italian?**
22 **A**: so did it work?
23 **What about people and Italian?**
24 **B: what worked? I'm losing the thread**'

In (8), lines 1–4, 6, 8 and 11 present an MDU. It starts with an absolutely free remark that introduces a new topic (*a ja proverjaju zadachku*) and ends with an absolutely bounded remark (an emoticon). However, this MDU is interrupted by another one in the middle (*ty mhe luchshe skazhi, s chem ty nosish' svoj seryj kardigan iz frantsii?*). The "cardigan topic" is presented in lines 5, 7, 9–10, 12–17. In (8), we haven't provided all the lines of the second MDU, as it is quite long; nevertheless, it can be noted that line 5 introduces an absolutely free remark and that all conditions of illocutionary force are satisfied. We can state, therefore, that in (8), two MDU are interlaced, both of them being important and "intentional". Same can be said about (9). Here, lines 4, 8, 18, 20 present an incorporated MDU that both communicants are eager to develop.

In our data, all long dialogues and 9 medium dialogues present interlaced MDU. However, only in two cases it leads to communicative failure. Characteristics of these dialogues suggest the hypothesis (11).

(11) If an Internet dialogue has all the characteristics of an oral dialogue except the communication channel (both communicants have one and only one interlocutor; they are not involved in any other mental activities; they talk without pauses in a "close-to-oral-speech" rhythm), an overlooking of remarks and recursion obligatorily occur.

The reason for that, as was indicated earlier, is lack of simultaneity between producing and perceiving of a remark. A communicant, eager to develop a conversation or to introduce a new topic, is unable to register all the remarks of his interlocutor as soon as they appear on the screen. As we have seen earlier, the result of such overlooking are two interlaced intentional MDU, a phenomenon that is not possible in oral speech.

Nevertheless, the situation described above is not so frequent and almost never leads to a communicative failure. That is because instant messaging is very rarely used as an exact analog of oral commination. Rather, it is a completely different type of communication, where multitasking is a fundamental principle. Normally, when people talk via Internet, they introduce a remark and turn to some other activities, thus providing a pause needed for the response.

Conclusion

In this article, a correlation between oral and Internet dialogue have been analyzed. It has been noted that comparison of these two types of dialogue is rarely appropriate, because they differ not only in the communication channel, but also in some other ways. In spite of that, some Internet dialogues that seem to satisfy "close-to-oral-conversation" conditions have been analyzed. These dialogues demonstrate a feature that is impossible in oral communication, that is, an interlacing of two intentional MDU. If the communicants maintain a "close-to-oral-conversation" rhythm, the interlacing results in communicative failure. This fact demonstrates that instant messaging doesn't work as an analogue of oral communication, but rather as a new medium that favors multitasking.

References

1. *Baranov A. N., Kreidlin G. E.* (1992), Illocutionary force in the dialogue structure [Illokutivnoe vynuzhdenie v strukture dialoga], Voprosy yazykonania [Problems of linguistics] no. 2, pp. 84–99.
2. *Baron N. S.* (2010) "Discourse structure in instant messaging: the case of utterance breaks", available at: http://www.languageatinternet.org/articles/2010/2651

3. *Herring, S. C.* (1999), "Interactional coherence in CMC", available at: http://jcmc.indiana.edu/vol4/issue4/herring.html
4. *Berglund T. Ö.* (2009) "Disrupted turn adjacency and coherence maintenance in instant messaging conversations", available at: http://www.languageatinternet.org/articles/2009/2106
5. *Voida A., Newstetter W. C., Mynatt E. D.* I think, therefore IM: When conventions collide: the tensions of instant messaging attributed. Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves. New York, NY, 2002, pp. 187–194.
6. *Lam Ch., Mackiewitcz J.* A case study of Coherence in a workplace instant messaging. Professional Communication Conference. Seattle, WA, 2007, pp. 1–6.
7. *Sacks, H., Schegloff, E. A., Jefferson, G.* (1974) A simplest systematics for the organization of turn-taking for conversation. Language, Vol. 50, pp. 696–735.
8. *Traum, D. R.* (1999), Speech acts for dialogue agent, in Foundations Of Rational Agentcy, Kluwer Academic Publishers, Dordrecht pp. 169–201.
9. *Zemskaia E. A., Kitaigorodskaia M. V., Rozanova N. N.* Russian oral speech [Russkaia razgovornaya rech']. Nauka, Moskva, 1983.

О СТРУКТУРЕ ТЕЗАУРУСА ПОСЛОВИЦ

Малкова А. С. (asmalkova@gmail.com)

Институт лингвистики РГГУ, Москва, Россия

С переходом к электронным способам хранения информации и появлением полнотекстового поиска одна из наиболее обсуждаемых проблем в паремиологии — проблема упорядочения словарных статей в сборниках пословиц — перешла в разряд чисто технических задач. Однако споры о том, как должен быть упорядочен пословичный материал (по алфавиту, по темам, по ключевым словам), нужен ли индекс, если да, то с какой степенью детализации — в действительности имеют под собой существенное методологическое основание, а именно: каким образом те способы поиска и навигации, которые впоследствии окажутся в распоряжении пользователя, закладываются на этапе планирования структуры словаря. Уже сама структура словаря может как усложнить процесс поиска, так и облегчить его.

Появление компьютерных систем решает задачу текстового поиска, открывая тем самым дорогу для новых вызовов, главный из которых — поиск семантический. В отношении последнего все вышесказанное также справедливо. Структура семантического аппарата существенно влияет на то, насколько поисковая система будет удобной в эксплуатации. Причем не только для конечного пользователя, но и для составителя словаря, и его модераторов.

В настоящей статье предлагается подход к построению формального аппарата тезауруса пословиц, при котором паремиологический фонд представляется в виде динамически генерируемой семантической сети, т. е. семантические связи между пословицами конструируются автоматически в процессе работы пользователя с системой.

Предложенный подход был реализован в экспериментальной системе МЕТАФОРА (система находится в открытом доступе по адресу http://metaphora2.ru), послужившей источником иллюстративного материала для статьи.

Ключевые слова: тезаурус, идеографический словарь, пословица, структурная паремиология

Динамически генерируемая семантическая сеть

Семантическая сеть — это множество узлов, связанных между собой дугами. Формальное описание сети возможно не только через явное перечисление всех ее элементов: дуг и узлов. Если узлы не атомарны, а представляют собой составные объекты, дуги могут быть заданы как система отношений на множестве узлов.

Поясним сказанное на простом примере: пусть у нас имеется коллекция предметов мебели, где каждый предмет описан набором характеристик (высота, количество ножек, назначение и т. п.). Тогда введение отношения,

в котором параметр «количество ножек» имеет фиксированное значение (например, четыре), равносильно попарному связыванию дугами всех предметов с четырьмя ножками. Построение сети через описание отношений кажется нам наиболее адекватным решением при проектировании тезауруса пословиц. Почему?

В паремиологии сейчас утвердилось мнение, что пословицы являются знаками ситуаций, в контексте которых они употребляются [Пермяков Г. Л. 1979, Жолковский А. К. 1978]. Этим объясняется, почему к пословичному фонду слабо применимы семантические отношения, использующиеся в лексикографии. Здесь, также как и во фразеологии, даже синонимия никогда не бывает полной [Баранов А. Н., Добровольский Д. О. 2010]. Семантические отношения должны базироваться на сходстве структуры моделируемых ситуаций.

При этом одно высказывание может входить в различные классы подобия в зависимости от того, на каком аспекте ситуации акцентируется внимание. Например, пословица *Кто рано встает, тому Бог дает.* сходна

- по линии *трудолюбие* с текстом *Под лежачий камень вода не течет.*
- по линии *заслуженная награда* с *По труду и честь.*
- по линии *личная инициатива* с *Человек — кузнец своего счастья.*
- по линии *оперативность в действиях* с *Куй железо, пока горячо.*
- по линии *упорство в действиях* с *Капля камень точит.*

и так далее. Красивую метафору, описывающую это явление, мы находим в статье Ю. Левина: «пословичный фонд представляет собой <…> не просто систему единиц <…>, но и некоторое «пространство» в математическом смысле слова» [Левин Ю. И. 1984], на котором задана «метрика» — отношение семантического сходства.

Очевидно, что явно формировать такую сеть перечислением всех дуг, поддерживать ее в актуальном состоянии крайне сложно. Сложно как с методологической точки зрения, так и с чисто практической, принимая во внимание поистине колоссальное количество таких связей.

Решение, которое мы предлагаем в настоящей работе, заключается в разработке метаязыка, с помощью которого структура пословичной ситуации может быть описана формально, и создании на базе этого языка системы семантических отношений, связывающей пословицы в единую сеть.

Предпосылки создания метаязыка

Кажется, что разработать формальный язык, адекватно представляющий структуру пословичной ситуации, еще с учетом значений, возникающих в контекстах, крайне сложно.

Да, если рассматривать формальное описание как альтернативу толкованию пословицы. Однако при проектировании тезауруса задача ставится несколько иначе. Рассмотрим такие возможности практического применения тезауруса:

1) пользователь вводит понятие, например *трусость, скромность, одиночество, хвастовство*, с намерением найти тексты, в которых это понятие раскрывается;

2) пользователю необходимо подобрать более точное, «меткое» высказывание, конкретизировать свою мысль, и для этого перейти от некоторого заданного текста к семантически близким пословицам (но при этом не полностью синонимичным).

В обоих случаях система выдает пользователю список высказываний, из которых он выбирает наиболее подходящее. Наличие шума здесь не столь критично, существенно большей проблемой является недостаточная полнота выборки. Таким образом, ключевой оказывается не столько способность системы быть чувствительной к оттенкам семантики, сколько наоборот, ее способность находить общую часть в семантически нетождественных текстах.

Суть предлагаемого решения сводится к тому, чтобы выявить очень компактный набор достаточно абстрактных базовых понятий, таких как ТРУД, ВОЛЯ, УСПЕХ, ВЛАСТЬ. Их семантические поля должны почти не пересекаться. За счет комбинирования базовых понятий в составе метаописаний пословиц их семантика будет сильно сужаться, конкретизироваться. Например, базовое понятие ТРУД–БЕЗДЕЛЬЕ присутствует в метаописаниях пословиц, говорящих о *труде, целеустремленности, добросовестности, предусмотрительности, ответственности* и их антонимах *халатности, лени* и т. п. Понятие ВОЛЯ–ПОКОРНОСТЬ — в пословицах о *смелости и инициативе, установлении своих правил* и, напротив, о *подчинении чужим людям и обстоятельствам, о трусости и безволии* и т. д. Наличие общего базового понятия в метаописаниях различных пословиц является формальным основанием для признания их частичного сходства. При этом различия в оставшейся части метаописаний обеспечивают конкретизацию понятия.

При выявлении базовых понятия и структуры метаописаний, мы отталкивались от положений когнитивной теории метафоры Дж.Лакоффа [Lakoff G., Johnson. M 1980] и от концепции мифологического нарратива О. М. Фрейденберг [Фрейденберг О. М. 1997]. Подробно процесс выявления понятий изложен в [Малкова А. С. 2011].

Система Г. Л. Пермякова

Прежде чем мы перейдем к описанию предлагаемого метаязыка, нельзя не сказать два слова о системе Пермякова [Пермяков Г. Л. 1979], поскольку она как раз представляет собой попытку создания такого формального языка. Чтобы не перегружать повествование, ограничимся очень сжатым пересказом основных положений и перейдем к содержательной критике, также весьма краткой.

Модель Пермякова, по замыслу ее создателя, отражает чистую структуру ситуации, освобожденную от художественно-поэтических средств языка. Формальная модель любой пословицы представляет собой комбинацию двух

структурных элементов: 1) тематической пары[1] и 2) типа отношения между элементами пары[2].

На основе анализа обширного материала Пермяков формирует списки тематических пар и типов отношений. Тематических пар в системе Пермякова насчитывается 180, при этом 78 основных, охватывающих порядка 97% пословичного фонда. Количество типов отношений — 64.

Хотя система задумывалась Г. Л. Пермяковым как «язык», и сам автор называл ее «*грамматика* пословичной мудрости» (курсив мой — А.М.), де факто здесь предложен не язык, а несколько более изысканная классификация со свойственными классификациям недостатками:

1) Семантические связи между рубриками не заданы, тематические пары приводятся длинным списком, среди них есть как семантически созвучные, такие как *Труд–Безделье* и *Труд–Отдых* или *Уродство–Красота* и *Форма–Содержание*, так и очень далекие: *Два–Один, Друг–Враг, Тайное–Явное*. Встречаются даже случаи дублирования семантики в тематических парах и отношениях: одновременно существуют пара *Порождающее–Порожденное* и отношение *Порождение–Непорождение*.

2) Необходимость «втиснуть» всю семантику пословицы противопоставление двух элементов тематической пары, приводит к появлению сложных для восприятия, очень абстрактных и потому непригодных для поиска в обычном понимании этого слова именований пар и в особенности типов отношений. Так, например, сложно себе представить, чтобы пользователю потребовалось найти пословицы, характеризующиеся отношением *Проявление–Непроявление одной вещи в другой*.

3) Вопреки первому впечатлению, подкрепленному уверениями автора о высоком комбинаторном потенциале модели, списки тематических пар и отношений сильно взаимообусловлены, т. е., определенные тематические пары фактически могут сочетаться только с определенными типами отношений. Лишь весьма ограниченная часть из всех возможных комбинаций подкрепляется реальными текстами. Отношения сильно несбалансированы. По приблизительным оценкам из 64 видов отношений одна группа *Качественное соответствие–Несоответствие* включает в себя более 10% пословичных высказываний, в то время как ряд других видов отношений крайне низкочастотны.

Работа Г. Л. Пермякова, безусловно, представляет собой существенный шаг на пути к формализации семантики пословицы, в ней был высказан целый ряд ценных идей: 1) четкое разграничение плана выражения и плана содержания; 2) указание на пути регулярной вариативности[3]; 3) обоснование

---

[1] Примеры тематических пар: *Правда–Ложь, Начало–Конец, Молодость–Старость* и т. п.

[2] Примеры отношений: *Качественное соответствие–Несоответствие, Однородность–Неоднородность, Неотделимость–Отделимость* и т. п.

[3] Пермяков подметил любопытный факт, противоречащий традиционному представлению о пословице как о «народной мудрости». В любом пословичном фонде можно

значимости бинарных противопоставлений в семантике пословицы. Однако система Пермякова мало пригодна для использования в тезаурусе, поскольку пословицы в ней не связываются в сеть. Они разбиты на 180 × 64 = 11 520 кластеров, содержащих синонимичные тексты, но при этом почти никак не связанных между собой.

Элементы метаязыка

Основным элементом предлагаемого метаязыка является ***оппозиция*** — пара противопоставленных понятий. Все оппозиции можно разделить на два класса:

1) оппозиции, характеризующие ЧЕЛОВЕКА, его личностные качества и поступки:

- *добродетель–порок*: моральный аспект;
- *знание–глупость*: интеллектуальный аспект;
- *труд–безделье*: деятельностный аспект;
- *помощь–вред*: социальный аспект;
- *воля–покорность*: волевой аспект.

2) оппозиции, характеризующие внешние по отношению к человеку обстоятельства (МИР), возникающие как в результате поступков человека, так и независимо:

- *успех–неудача*: достижение целей;
- *доход–убыток*: обладание благами;
- *радость–мука*: моральное и физическое удовлетворение;
- *работа–покой*: частный случай оппозиции *радость–мука:* подневольный, неблагодарный труд vs. отдых, отсутствие забот;
- *дружба–одиночество*: социальное приятие (любовь и дружба);
- *власть/почет–презрение*: социальный статус.

Пара оппозиций, связанных отношением, образует элементарное высказывание, или ***звено***. В графической записи оппозиция изображается как две расположенные одна над другой окружности (верхняя окружность соответствует позитивной части оппозиции, а нижняя — негативной*)*. Звено изображается как две стоящие рядом оппозиции.

встретить высказывания, противоречащие друг другу:

У льва родится лев. и *Отец лев, а сын — шакал.*

Все течет, все изменяется. и *Все возвращается на круги своя.* и т. п.

Пермяков видел причину этого явления в иллюстративной функции пословиц, а их варьирование объяснял многообразием ситуаций, в контексте которых пословица может употребляться.

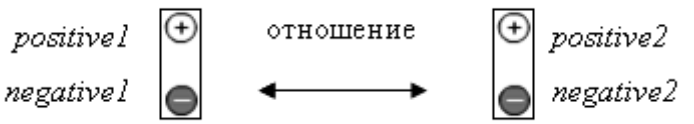


Рис. 1. Графическое изображение звена

Поскольку все оппозиции делятся на два класса (МИР и ЧЕЛОВЕК), возможно четыре варианта их сочетания в звене:

1. ЧЕЛОВЕК → МИР: влияние поступков, черт характера человека на его судьбу, отношение к нему других людей, например: *кто трудится, будет жить в достатке*.
2. МИР → ЧЕЛОВЕК: влияние внешних обстоятельств на поступки человека, необходимость совершать вынужденные действия, например: *что пропало, того уже не вернуть*.
3. ЧЕЛОВЕК < ЧЕЛОВЕК: установление приоритетов между чертами характера человека или внутри коллектива, например: *главное не внешность, а сущность человека*.
4. МИР < МИР: установление приоритетов между внешними обстоятельствами, например: *дружба важнее богатства*.

В высказываниях, где оппозиции принадлежат к разным классам, отношение может быть охарактеризовано как «влияние», где к одному — как «предпочтение».

Следующий элемент звена — это ***отождествление*** (семантический маркер отношения). Семантика отождествления может быть понятна из следующего примера. Рассмотрим пословицы:

отождествление / **оппозиция**	*внешность < сущность с виду positive, а на самом деле negative*	*временно < в итоге временно может быть negative, главное, чтобы в итоге positive*
добродетель–порок	В тихом омуте черти водятся.	Повинную голову меч не сечет.
знание–глупость	Вырос, а ума не вынес.	Век живи, век учись.

Пословицы в первой строке основаны на противопоставлении элементов оппозиции *добродетель-порок*, во второй — оппозиции *знание–глупость*. В столбцах присутствует противопоставление еще по одному признаку: *внешний вид–сущность* и *временное–итоговое* состояния.

Природа противопоставления элементов отождествления восходит к когнитивному механизму, исследованному и описанному

О. М. Фрейденберг [Фрейденберг О. М. 1997] на материале мифологии античности. Это бессознательное ощущение глубинной взаимосвязи между двумя явлениями вплоть до полного стирания границ между ними. Исследовательница указывает на основные направления отождествления — это равенства *целого* и *части*, *мнимого* и *реального*, *прошлого, настоящего* и *будущего*, *актера* и *роли*, *акции* и *реакции, постоянного* и *временного*, *лидера* и возглавляемого им *коллектива* и другие. Со временем связь в архаических отождествлениях ослабевает, однако они продолжают служить основой для фольклорных сюжетов.

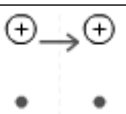
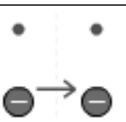
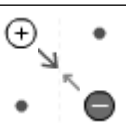
В метаязыке отождествление маркирует связь между оппозициями в звене. Первая оппозиция соотносится с первой частью отождествления, вторая — со второй. В метаязыке присутствуют следующие отождествления:

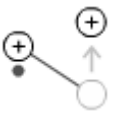
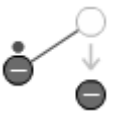
- *временно–в итоге*: противопоставление временны**х** состояний;
- *иногда–всегда*: другое п. временны**х** состояний;
- *мнимо–реально*: п. ложного (мнимого) и истинного;
- *желаемое–доступное*: п. модальностей ХОТЕТЬ и МОЧЬ;
- *материя–дух*: п. ценностей по духовной шкале;
- *часть–целое*: п. по количественной шкале;
- *некто–сам:* п. реального субъекта и его субститута.

Вторая часть во всех отождествлениях доминирует над первой (признается более значимой).

Последний элемент звена — это ***конфигурация***, регулярное сочетание ±-оценок входящих в звено оппозиций ***о1*** и ***о2*** с результирующей этической оценкой всего звена Σ (положительная результирующая оценка звена отражает идею одобрения или оправдания, отрицательная — идею несправедливости или осуждения). Конфигурация является независимым элементом звена и обладает собственной семантикой. Список конфигураций приведен в таблице 1.

Таблица 1. Графические схемы конфигураций

Оценки			Граф. схема		Пример интерпретации
о1	***о2***	**Σ**	**а)**	**б)**	
ТИП «ВЛИЯНИЕ»					
+1 –1	+1 –1	+1 +1			а) кто *трудится*, будет иметь *доход* б) кто *бездельничает*, потерпит *убыток*
+1 –1	–1 +1	–1 –1			а) некто *трудился*, а терпит *убытки* б) некто *бездельничал*, а имеет *доход*

Оценки			Граф. схема		Пример интерпретации
o1	*o2*	Σ	а)	б)	
ТИП «ПРЕДПОЧТЕНИЕ»					
–1 +1	+1 –1	+1 –1			а) пусть *часть* плох., зато *целое* хорошее б) хотя *часть* хорошая, но *целое* плохое
+1 –1	–1 +1	+1 –1			а) когда *целое* плох., хор. *часть* — уже хорошо б) плохая *часть* портит хорошее *целое*
+1 –1	+1 –1	+1 –1			а) *целое* хорошее и *часть* хороша б) *целое* плохое и *часть* плоха

Таким образом, элементарное высказывание (звено) состоит из четырех элементов: двух оппозиций, отождествления и конфигурации.

Помимо графической записи используется также сокращенная текстовая запись звена. В ней сферы (ЧЕЛОВЕК, МИР) выделяются заглавными литерами, отождествления — курсивом, слова, служащие для передачи конфигурации — полужирным, у каждой из двух оппозиций незадействованная часть опускается.

Рассмотрим два примера звеньев, их текстовые записи и графические представления:

Пример 1. *Временно человек ошибался (совершал дурные поступки), но в конце концов он исправился.*

ЧЕЛОВЕК[**хотя** *временно*(порок), **зато** *в итоге*(добродетель)].

Рис. 2. Графическая схема звена (пример 1)

Пример 2. *Кто трудится, будет жить в достатке.*
если ЧЕЛОВЕК[(труд)], **то** МИР[(доход)].

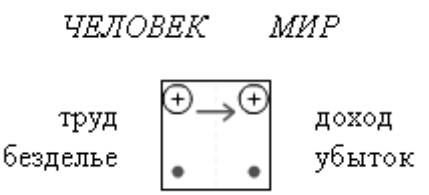


Рис. 3. Графическая схема звена (пример 2)

Полное метаописание пословицы включает от одного до трех звеньев: одно звено целиком лежит в сфере ЧЕЛОВЕК, другое — целиком в сфере МИР и еще одно звено объединяет обе сферы (связь направлена либо в одну, либо в другую сторону). Графически метаописание пословицы может быть представлено в виде следующей трехчастной структуры:

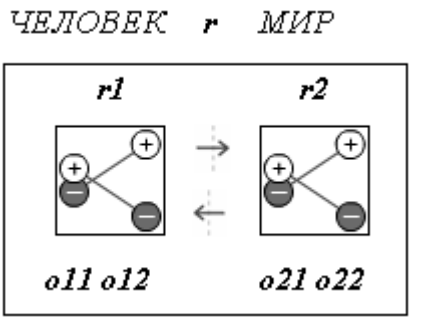


Рис. 4. Полная графическая схема пословицы

Экспериментальная выборка

Границы применения формального языка исследовались на экспериментальной выборке пословиц. Для целей исследования было необходимо, чтобы экспериментальная выборка была не слишком объемной, но при этом хорошо покрывала все основные темы, присутствующие в пословичном фонде. Такими свойствами обладает достаточно объемный верхний срез частотного словаря. В качестве экспериментальной выборки был использован словарь В. М. Мокиенко [Словарь русских пословиц 2007], содержащий 505 наиболее употребительных современных пословиц. Словарь хорош тем, что он почти в точности воспроизводит «паремиологический минимум» Г. Л. Пермякова (раздел первый *Пословицы*, подразделы I *Собственно пословицы* и II *Народные афоризмы*) и при этом содержит толкования и примеры употребления пословиц в художественной литературе. Несмотря на некоторые погрешности словаря (смешение в нем пословиц и поговорок; недостаточное отражение вариативности

в заголовках словарных статей и т. п.), он позволяет сформировать представление о наиболее частотных пословичных моделях. Бесспорно, в дальнейшем выборка должна быть расширена для уточнения элементов и конструкций метаязыка и расширения его выразительности.

В настоящий момент численные результаты, полученные на экспериментальной выборке, следующие:

– около 70 % пословиц хорошо представимы с помощью метаязыка: формальное описание достаточно адекватно передает основной смысл пословицы;
– около 20 % пословиц представимы частично: формальное описание отражает один аспект семантики пословицы или отражает ее неточно;
– еще около 10 % пословиц не могут быть адекватно представлены; это пословицы с очень узкой или специфической семантикой.

Семантический поиск с использованием метаязыка

Как было сказано выше, использование метаязыка для описания структуры пословичной ситуации позволяет динамически генерировать семантическую сеть пословиц. В настоящем разделе будет показано, как конструируется эта сеть.

Пусть пользователь выбрал в базе данных некоторый текст (назовем его *эталонным*). Тогда сразу же относительно этого текста с помощью стандартных процедур могут быть автоматически найдены: 1) подобные (почти синонимичные) пословицы; 2) противоположные (антонимичные[4] либо контрапозитивные[5]) тексты; 3) частично сходные.

Приведем конкретном примере. В качестве ***эталонного*** текста возьмем пословицу *Смелость города берет*. В базе данный ей сопоставлено метаописание:

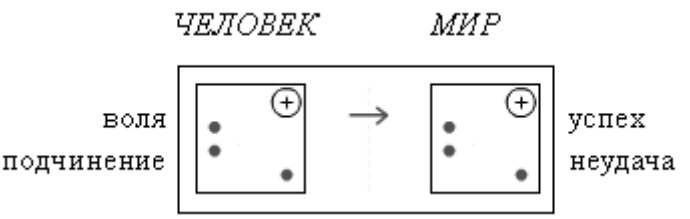


Рис. 5. Графическая схема эталонной пословицы

т. е. ‘проявление воли приводит к успеху’.

Подобные пословицы обладают в точности таким же метаописанием:

---

[4] *Антонимия* здесь понимается как описание противоположных ситуаций (*внешность хорошая, натура плохая* и наоборот *внешность плохая, натура хорошая*).

[5] *Контрапозиция* понимается как выражение одной и той же идеи через отрицание (*знание приводит к успеху, глупость приводит к неудаче*).

Кто смел, тот и съел.
Риск — благородное дело.

Противоположные (в данном случае, *контрапозитивные*) пословицы содержат те же отождествления и оппозиции, но противоположные результирующие оценки конфигурации:

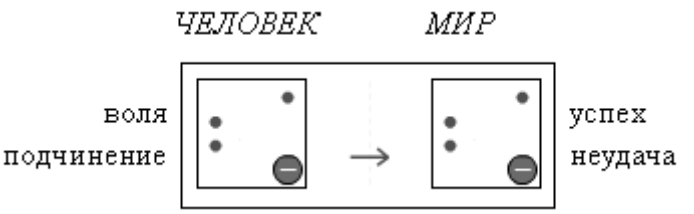


Рис. 6. Графическая схема противоположных пословиц

т. е. ‘проявление слабохарактерности приводит к поражению’.

Волков бояться — в лес не ходить.
Кто не рискует, тот не пьет шампанское.

Частично сходные пословицы — это те, у которых совпадают два элемента в любом из непустых звеньев: доминирующая оппозиция + отождествление, или доминирующая оппозиция + конфигурация, или отождествление + конфигурация. Это одновременно простой и достаточно сильный критерий для того, чтобы в семантике пословицы носитель языка ощущал существенное сходство.

а)

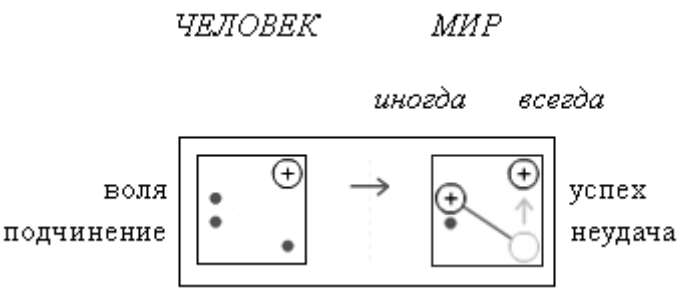


Рис. 7. Графическая схема частично сходных пословиц (а)

т. е. ‘нужно проявлять волю и надеяться на лучшее, авось повезет’.

Бог не выдаст, свинья не съест.
Двум смертям не бывать.
Раньше смерти не умрешь.

б)

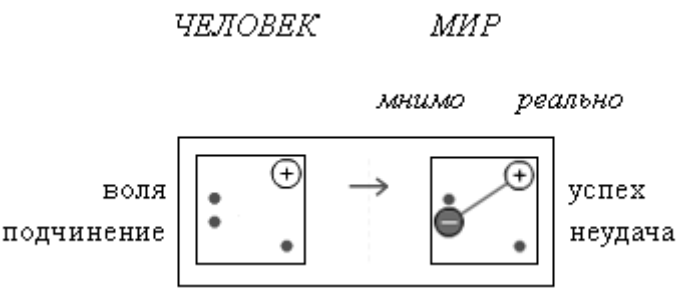


Рис. 8. Графическая схема частично сходных пословиц (б)

т. е. ‘нужно проявлять волю, опасность и неудача кажущиеся’.

Не так страшен черт, как его малюют.
Не боги горшки обжигают.

в)

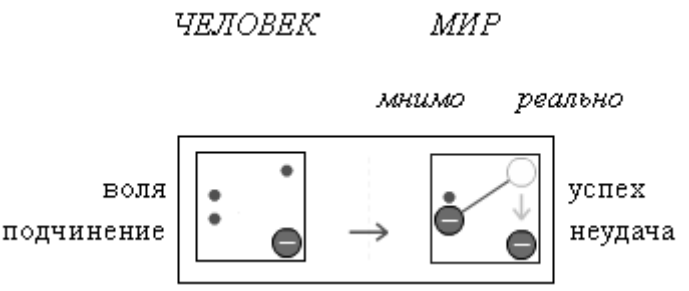


Рис. 9. Графическая схема частично сходных пословиц (в)

г)

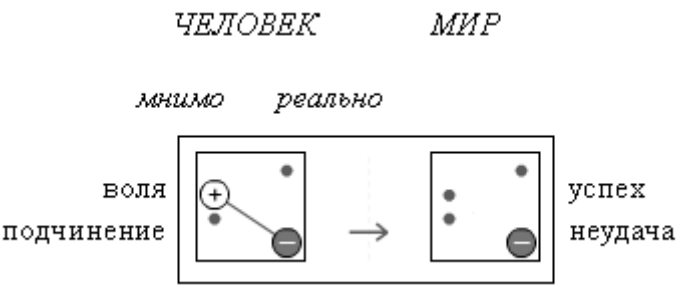


Рис. 10. Графическая схема частично сходных пословиц (г)

т. е. ‘проявление слабохарактерности приводит к тому, что мнимая неудача оборачивается реальной’.

У страха глаза велики.
Пуганая ворона куста боится.
Трус умирает тысячу раз.

т. е. ‘мнимая смелость при отсутствии сильного характера приводит к неудаче’.

Молодец среди овец.
Хвалилась синица море зажечь.

Метаязык может быть использован для решения еще одной задачи, а именно для поиска пословиц по отраженным в них ***понятиям***, таким как *терпение, трусость, опытность* и т. п. При традиционном подходе для этого требуется перечислить для каждой пословицы связанные с ней понятия (установить соответствие «каждый с каждым»). При использовании метаязыка достаточно каждому понятию поставить в соответствие фрагмент метаописания. Тогда соотнесение понятий и пословиц будет осуществляться автоматически исходя из наличия в метаописаниях пословиц соответствующих фрагментов.

Приведем ряд примеров (к сожалению, формат статьи не позволяет вместе с каждым понятием привести его пословицы; с примерами можно ознакомиться на сайте http://metaphora2.ru/ в разделе «Поиск по ключевым словам»):

учение: ЧЕЛОВЕК[**хотя** *временно*(глупость), **зато** *в итоге*(знания)]
предусмотрительность: МИР[**хотя** *временно*(работа), **зато** *в итоге*(покой)]
алчность: МИР[*желаемое*(доход), **а** *доступное*(убыток)]
смирение с потерями: **если** МИР[(убыток)], **то** ЧЕЛОВЕК[(покорность)]
терпение: МИР[**хотя** *временно*(неудача), **зато** *в итоге*(успех)]
злонамеренность **если** ЧЕЛОВЕК [(вред)], **то** МИР[(*negative*)]
самопожертвование: МИР[**хотя** *материя*(страдание), **зато** *дух*(дружба)]
везение: МИР[*иногда*(успех) **компенсирует** *всегда*(неудача)]
опыт: **если** ЧЕЛОВЕК [(знание)], **то** МИР[(успех)]
неблагодарность: **несправедливо, что** ЧЕЛОВЕК [(помощь)], **а** МИР[(*negative*)]
обещание: ЧЕЛОВЕК [*мнимо*(*positive*) **и** *реально*(*positive*)],

В настоящий момент в базу системы МЕТАФОРА введено около ста несинонимичных понятий, среди которых: *богатство, неисправимость, правда, сила/власть, основательность, невзгоды, кара, дружба, закон, внешность, беспечность, своеволие, жадность, несвоевременность, промахи, спешка, одиночество, соблазн, отчаянность, трусость, недостойный, судьба, бестолковость, неопытность, скромность, ссора, тайна, помощь, риск* и другие. Новые понятия заносятся в базу и выверяются в итеративном процессе анализа пословиц из выборки.

Заключение

В статье излагаются основы метаязыка, моделирующего структуру пословичной ситуации (описываются словарь, основные конструкции). Множество полученных метаописаний обладает сетевой структурой и, таким образом, может послужить ядром тезауруса пословичных изречений.

Применение метаязыка может быть востребовано в перспективе решения двух задач: 1) поиска семантически сходных (но при этом несинонимичных) высказываний; 2) поиска пословиц по затронутым в них понятиям.

Положения статьи проиллюстрированы примерами из экспериментальной информационной системы МЕТАФОРА, доступной по адресу http://metaphora2.ru/.

Литература

1. *Lakoff G., Johnson. M.* (1980). Metaphors We Live By. Chicago, University of Chicago Press, 1980.
2. *Баранов А. Н., Добровольский Д. О.* (2010) Семантика фразеологизмов: иерархия или сеть? // «Компьютерная лингвистика и интеллектуальные технологии»: Труды Международной конференции «Диалог 2010». Бекасово, 2010. — С. 25–29.
3. *Жолковский А. К.* (1978), К описанию выразительной структуры паремий. // Паремиологический сборник. Пословица. Загадка. п/р Г. Л. Пермякова. — М.: «Наука», 1978.
4. *Левин Ю. И.* (1984), Провербиальное пространство. // Паремиологические исследования. п/р Г. Л. Пермякова. — М.: «Наука», 1984.
5. *Малкова А. С.* (2011), Методика создания словаря базовых концептов для формальной модели семантики пословицы // Материалы XL Международной филологической конференции 14–19 марта 2011 г. Психолингвистика / Отв. редактор Т. В. Черниговская. — СПб.: Филологический факультет СПбГУ, 2011. — С. 18–21.
6. *Пермяков Г. Л.* (1979), Пословицы и поговорки народов Востока. Систематизированное собрание изречений двухсот народов. — М.: 1979.
7. *Словарь* русских пословиц (2007), В. М. Мокиенко, Ю. А. Ермолаева, А. А. Зайнульдинов и др.; под ред. В. М. Мокиенко. — М.: Астрель: АСТ, 2007.
8. *Фрейденберг О. М.* (1997), Поэтика сюжета и жанра. — М.:Лабиринт, 1997.

РАЗРАБОТКА ЛЕКСИКО-СЕМАНТИЧЕСКОГО СЛОВАРЯ КИТАЙСКОГО ЯЗЫКА ДЛЯ МНОГОЯЗЫЧНОЙ СИСТЕМЫ АНАЛИЗА ТЕКСТОВ

Маничева Е. С. (Ekaterina_M@abbyy.com),
Дрейзис Ю. А. (Yulya_D@abbyy.com),
Селегей В. П. (Vladimir_S@abbyy.com)

ABBYY, Москва, Россия

В статье описывается продолжающийся проект создания лексико-семантического словаря ядерной лексики китайского языка для системы автоматического анализа естественного языка, разрабатываемой в компании ABBYY. Оцениваются особенности задачи описания китайского языка в многоязычном проекте, ориентированном на универсальные семантические представления. Приводятся некоторые методологические трудности, с которыми столкнулись разработчики и принятые решения. Созданный лексико-семантический ресурс может представлять ценность как для систем автоматического анализа текстов, так и как основа учебного словаря китайского языка, адресованного прежде всего обучаемым с повышенными требованиями к полноте и системности лексикографического описания.

Ключевые слова: лексико-семантический словарь, компьютерная лексикография, китайский язык, глубинная семантическая модель, лексико-грамматическая сочетаемость

Введение. Китайский язык — проверка универсальности моделей описания

Для многоязычных систем компьютерного анализа текстов, основанных на лингвистических моделях (model-based), особую важность имеет проверка адекватности используемых моделей языковых описаний для типологически сильно различающихся языков.

Проект ABBYY Compreno [1] ориентирован на создание различных технологий анализа текста на основе универсальных языковых моделей. По многим соображениям работы по созданию языковых описаний начались с русского и английского языков. В ходе этой масштабной работы модель данных претерпела существенные изменения. Расширение списка рабочих языков за счет французского и немецкого языков не позволяло говорить о реальной типологической корректности модели описания.

В этом смысле начатая в 2008 г. работа над китайским описанием явилась настоящей проверкой проектных принципов. В данном докладе речь пойдет о тех задачах и проблемах, которые связаны прежде всего с лексико-семантическим описанием. Как кажется авторам, и проблемы, с которыми мы столкнулись, и созданный лексикографический ресурс имеют значение, выходящее за рамки отдельного, пусть даже и весьма амбициозного, проекта.

1. Подготовительный этап: ресурсы и методики

Перед началом работ по составлению словаря был осуществлен мониторинг центров разработки, технологий и ресурсов по автоматической обработке китайского языка, китайской системной лексикографии, компьютерной и корпусной лингвистике, а также проектов по разработке моноязычных и многоязычных лексико-семантических баз.

На начальном этапе проекта нам была очень важна заинтересованная помощь коллег из РГГУ, ИССА, СПбГУ. В ходе совместных встреч и семинаров были обсуждены многие вопросы, легшие в основу методики описаний[1]. Поскольку имеющиеся лексические ресурсы для китайского языка оказались более скромными с точки зрения полноты лексикографического описания, чем ресурсы для других языков проекта, было принято решение, что лексико-семантическая часть китайского проекта должна быть относительно независимой, ориентированной не только на задачи автоматического анализа в Compreno, но и на создание учебного словаря китайского языка. Это обусловило интерес к проекту китаистов, не являющихся участниками проекта Compreno.

1.1. Ситуация с ресурсами по автоматической обработке китайского языка

Анализ имеющихся ресурсов по автоматической обработке текста (далее АОТ) показал, что лексико-семантической основой для решения различных задач по автоматической обработке китайского языка в Китае и на Тайване используются семантические сети, построенные по принципу WorldNet [3, 6, 7, 10,11]. При этом все они, за исключением разрабатываемого под руководством Центра компьютерной лингвистики Пекинского Университета (КНР) словаря понятий китайского языка Chinese Concept Dictionary (CCD), идут от структуры английского WordNet'a, а не китайского языка [7, 10].

Переводной принцип неизбежно приводит к существенной неполноте и искажениям в системе лексических значений языка. Кроме того, существующие

---

[1] Особенно ценным был вклад Тараса Ивченко, Юрия Музыченко, профессора Пекинского Университета Юй Шивэня, Веры Барас, Ирины Горбуновой, которым мы приносим искреннюю благодарность за участие в обсуждении методик на разных этапах проекта.

лексические ресурсы оказались чрезвычайно бедными с точки зрения полноты лингвистической информации о содержащихся в них лексических единицах.

Таким образом, стало очевидным следующее:

- необходимо выбрать базовое лексическое описание китайского языка, хорошо покрывающее его частотное ядро;
- необходимо верифицировать систему лексических значений с точки зрения модели и дистрибуции и дополнить описание всеми необходимыми типами описаний;
- необходимо провести работу по отображению описываемых значений на структуру семантических классов универсальной лексико-семантической иерархии, в рамках которой происходит описание языков в Compreno.

1.2. Разработка методики лексикографического портретирования

В качестве базового словаря ядерной лексики было решено использовать китайский словарь HSK 汉语8000词词典A Dictionary of Chinese Usage [12]. Данный словарь был разработан в КНР в качестве лексического минимума для иностранцев, сдающих экзамен на определение уровня владения китайским языком. Словарь HSK8000 включает в себя 8822 словарных входа, упорядоченных по 4 приоритетам частотности. К первому приоритету относятся 1033 словарные статьи, 2018 — ко второму, 2202 — третьему и 3569 словарных статей — к четвертому приоритету. По мнению составителей словаря, этот лексический запас покрывает 95 % современной лексики.

Разработка лексико-семантического словаря конкретного языка основывается на универсальной модели, разработанной в рамках проекта Compreno [1]. По технологии работы собственно лингвистическому этапу описаний предшествует этап подготовки лексикографического описания, результатом которого становится словарь, который можно рассматривать как относительно независимый промежуточный результат описания.

В основных положениях лексикографическое описание строится на принципах, близких к принципам системной лексикографии, сформулированным Ю. Д. Апресяном в [2]. Описание должно в явном виде содержать все типы лексикографической информации, которые необходимы для процедуры включения лексического значения в лексико-семантическую иерархию, используемую в технологиях Compreno.

Для китайского язык пришлось существенным образом уточнить методику лексикографического описания. В начале работ были выделены 3 основных лексикографических типа статей (для глагола, прилагательного и существительного) и определены структуры описания лексики каждого из типов.

Лексикографическая статья должна была содержать максимум лексико-семантической и синтаксической информации, релевантной для будущих задач автоматической обработки текста. Для всех лексикографических типов указывалась следующая информация: транскрипция, написание в полной форме, орфографические варианты, перевод на русский и английский языки,

толкование, семантическая и синтаксическая модель, сочетаемость с грамматическими модификаторами, возможность употребления в различных синтаксических позициях, стандартная сочетаемость, идиоматика, синонимы и антонимы. При описании глаголов и прилагательных необходимо было проиллюстрировать примерами глубинную и поверхностную модель, указать возможные залоговые трансформации, возможность образования редуплицированных форм и случаи именного употребления.

В процессе подготовки был разработан пакет методических документов для лексикографов: методика подготовки лексикографического портрета, инструкция по работе с лингвистическими ресурсами, указания по глоссированию примеров, расширенный список семантических ролей (называемых в проектной терминологии глубинными позициями, или ГП) с примерами их выражения в китайском языке. В качестве базовых источников по синтаксическому употреблению были выбраны 3 словаря грамматической сочетаемости: 1) HSK GTU: HSK词语用法详解. A Guide to the Usage of HSK Vocabulary [15]; 2) DC-CC-YF: 汉语动词用法词典 (Словарь употребления глаголов) [13]; 3) XRC-CC-YF: 汉语形容词用法词典 (Словарь употребления прилагательных) [14]. Недостающую или недостоверную информацию лексикографы должны были проверять по другим печатным и электронным словарям, корпусам, конкордансам и иным надежным источникам в литературе и интернете.

2. Интеграция китайского языка в проектную идеологию

Поскольку основной задачей доклада является прежде всего описание наиболее модельно-независимой компоненты — собственно лексикографического описания, мы не будем касаться здесь деталей устройства иерархии, в которую, в конечном счете, оно отображается. Остановимся только на тех элементах структуры иерархии, которые явно присутствуют в исходном лексикографическом описании и могут рассматриваться как объективные свойства языка, а не артефакты реализации.

Семантическая иерархия представляет собой лексико-семантический словарь, в котором содержится вся лексика языка, необходимая для анализа и синтеза текста в процессе перевода. СИ организована как дерево родо-видовых отношений, в узлах которого находятся **семантические классы** (СК) — универсальные (единые для всех языков) объекты, отражающие некоторое понятийное содержание, и **лексические классы** (ЛК) — конкретноязыковые объекты. СК характеризуются единичным наследованием и моделью семантических отношений (глубинной моделью), которая отражает потенциально возможные отношения с другими классами в любом естественном языке.

Глубинная модель описывается через глубинные позиции (ГП) — семантические отношения, в которые вступает управляющий элемент со своими зависимыми. Лексические потомки СК в конкретных языках описываются в иерархии через взаимосвязь глубинной модели управления и поверхностных форм выражения этих отношений, или диатезные соответствия.

В одном СК могут объединяться синонимы, антонимы и регулярные дериваты, образующие общее семантическое поле. Например, в СК ‘TEMPERATURE’ в русском языке в качестве потомков расположены лексические классы *«холодный», «горячий», «тепло», «мороз», «греть», «разогревать»* и ряд других. Все ЛК отмечены смыслоразличительными атрибутами — **семантемами,** являющимися элементарными семантическими признаками. Например, лексические потомки одного СК могут различаться по семантической категории **полярности** (*холодный/горячий, тепло/холод*), **локализации в пространстве** (*сзади/спереди/вокруг; выбежать/вбежать*), **размеру** (*дом/домище/домишко*), **стилистике** (*видеть/зреть*) и многим другим.

Особенности синтаксического употребления и грамматического значения маркируются в СИ при помощи **граммем**. Например, глагол «зреть» отличается от глагола «видеть» не только стилевой окраской, но и грамматическими свойствами. Глагол «видеть» может подключать клоз при помощи союза «что», а его синоним «зреть» — нет. Данные особенности словоупотребления будут помечены в DPS соответствующей граммемой.

В соответствии с логикой СИ были проанализированы и критически пересмотрены принципы выделения значений исходного словаря HSK8000. Кроме того, в процессе заведения в СИ китайских классов стала формироваться система грамматических категорий, опирающаяся не столько на существующие грамматические теории, сколько на эмпирический материал, и вместе с тем согласующаяся с проектными принципами описания синтаксиса.

2.1. Критический взгляд на базовый словарь HSK8000

Словарь HSK8000 достаточно полно покрывает лексическое ядро и дает базовую информацию о грамматической сочетаемости описываемых единиц, именно поэтому он был выбран в качестве словарной базы и основы описания. Однако в ходе системных работ над лексико-семантическим словарем для системы АОТ стала наглядно видна его специфика как учебного словаря, ориентированного в первую очередь на иностранного студента. Главная задача словаря HSK8000 состоит не в том, чтобы системно описать лексику, а в том, чтобы помочь иностранцу овладеть базовым словарем и грамматикой. К несистемности описания лексики можно отнести следующие случаи:

1. Спорные приоритеты частотности.

Например, в 1 первый приоритет частотности попало слово liú xué shēng 留学生 *студент, проходящий стажировку за рубежом*, а на самом деле это слово частотно только в ограниченной языковой среде.

2. Непоследовательность в выделении единиц для словарного входа.

Словарными входами в словаре бывают единицы разных уровней:

- морфемы (化 huà в значении суффикса *-изировать; -изация*, 主义 zhǔ yì в значении суффикса *-изм*);
- слова (发展 fā zhǎn *развивать, развитие*, 父母 fù mǔ *родители*);
- словоформы с грамматическим показателем (养成 yǎng chéng *воспитывать* + показатель результатива);
- идиоматические единицы (一帆顺风 yì fān shùn fēng *идти как по маслу; без сучка, без задоринки*, 十全十美 shí quán shí měi *верх совершенства*, букв. 10 совершенств, 10 достоинств);
- свободные словосочетания (红旗 hóng qí *красный флаг*, 按劳分配 àn láo fēn pèi *распределять, распределение по труду)*;
- синтаксические показатели (着 zhe *показатель прогрессива,* 勿 wù *не, отрицание при императиве*);
- синтаксические конструкции (从。。。看来 cóng kàn lái *отсюда видно, можно сделать вывод*, 东跑西走 dōng pǎo xī zǒu *носиться туда-сюда*).

Для удобства коллективной работы словарь HSK8000 был поставлен под специальное программное приложение, в котором можно проводить полный мониторинг обработки значения от этапа назначения исполнителя до этапа проверки приписывания диатезных соответствий и делать SQL запросы.

По результатам анализа исходной словарной базы на первом этапе работ было решено отложить описание грамматических значений, синтаксических конструкций (670 значений) и идиоматических единиц (269 значений). Кроме того, из словаря HSK 8000 были исключены 3000 значений, которые не относились к базовым единицам словаря, а именно:

1. Свободные словосочетания, которые можно разбирать и переводить пословно.

Например:

Словосочетание 红旗 hóng qí *красный флаг*, предложная группа 按劳分配 àn láo fēn pèi *распределять, распределение по труду.*

В случае отсутствия в словаре вокабулы и соответствующего значения, лексикограф добавлял и описывал данное значение, а не значение всего словосочетания.

2. Словоформы с грамматическим показателем.

Например:

Словарный вход 养成 yǎng chéng в значении «воспитывать» был исключен в пользу аналогичного значения глагола 养 yǎng. Результативный показатель 成 chéng не является обязательным для выражения данного значения, оно может быть выражено как при помощи других результативных показателей, так и без них.

3. Случаи повторного выделения значения для слов в нехарактерной синтаксической позиции

Например: 发展 fā zhǎn *развивать* и *развитие*.

2.2. Критерии выделения значений в лексико-семантическом словаре

Типологической особенностью китайского языка является бедная морфология. Поэтому основным критерием для частеречной классификации в китайской лингвистической традиции становится критерий наиболее частотного синтаксического употребления слова в данном значении [8]. Большинство китайских двусложных предикатов может употребляться в именной функции как отглагольное имя, имя-характеристика, имя, обозначающее абстрактное понятие и имя, обозначающее конкретный объект. Разные случаи грамматической конверсии могут трактоваться как слова одной или разных частей речи в одном или нескольких значениях. В словаре HSK8000 критерии выделения значений и определения части речи не были эксплицитно сформулированы, и логика составителей не всегда представлялась последовательной.

Согласно проектным принципам описания языков, вместо понятия часть речи было решено оперировать понятием тип синтаксической парадигмы (ТСП), который описывает всё множество возможных составляющих с определенным типом элементов в качестве ядра. Следуя этим принципам, а также традициям отечественного [5] и западного [4, 9] китаеведения мы выделили следующие типы синтаксических парадигм для знаменательных слов: имя, предикат, наречие, числительное и местоимение. Для обозначения лексикализованных грамматических свойств слова в каждом конкретном значении использовались граммемы соответствующих грамматических категорий.

Проиллюстрируем разницу подходов к выделению значений и их грамматического типа в словаре HSK8000 и китайском словаре, построенном на базе СИ, на примере трех вокабул:

1. 发展 (fā zhǎn) : 发展经济 (fā zhǎn jīng jì) *развивать экономику*, 经济发展 (jīng jì fā zhǎn) — *развитие экономики*.

HSK8000: два значения для глагола *развивать* и существительного *развитие*

Китайский словарь в СИ (далее СИКит): одно значение, предикативный ТСП, возможность именного употребления обозначена граммемой.

2. 稳定 (wěn dìng) : 稳定关系 (wěn dìng guān xì) *стабилизировать отношения* 稳定的关系 (wěn dìng de guān xì) *стабильные отношения* 关系的稳定 (guān xì de wěn dìng) — *стабилизирование (стабильность) отношений*

HSK8000: одно значение для прилагательного *стабильный*

СИКит: одно значение, предикативный тип СП, возможность именного употребления, атрибутивного и именного употребления со значением характеристики промаркированы соответствующими граммемами и семантемой.

3. 翻译 (fān yì) : 翻译文章 (fān yì wén zhāng) — *переводить статью*, 文章的翻译 (wén zhāng de fān yì) — *перевод статьи*, 一篇翻译登在网站 (yī piān fān yì dēng zài wǎng zhàn) *опубликовать статью на сайте*, 一位翻译 (yī wèi fān yì) *один переводчик*.

HSK8000: два значения для глагола (*переводить*) и существительного (*переводчик*).

СИКит: три значения. Первое значение — *переводить* (предикат), именное употребление отмечено граммемой. Второе значение — *переводчик* (имя). Третье значение *перевод* как материальный носитель, тип текста (имя).

Таким образом, в соответствии с задачами АОТ главным критерием выделения значения является различие семантической и синтаксической модели, а тонкие различия в семантике и переводе до определенной степени вторичны. В том случае, если словарь будет использоваться для задач машинного перевода, точность перевода слова в том или ином языковом контексте можно будет обеспечить механизмами лексикализованного перевода при помощи переводных правил и словосочетаний. На данном этапе работ лексикограф только приводит примеры, не вписывающиеся в предлагаемый им переводной эквивалент, и по возможности указывает контексты, в которых предпочтителен иной перевод.

Например, у глагола 喂 wèi в словаре HSK 8000 выделено 2 значения: 1) кормить, давать еду животным и 2) давать еду, лекарство людям. Семантическая и синтаксическая модель обоих значений не имеет принципиальных различий. И в русском, и в английском языках перевод будет одинаковым независимо от того, дают пищу животным или людям,– «кормить кого-то чем-то, feed somebody with something». Однако если заполнителем данной глубинной позиции будет потомок СК, обозначающий лекарственные средства, предпочтительным представляется перевод «давать кому-то что-то». Если заполнителем глубинной позиции будет жидкий объект (имеющий семантему <<Liquid>>), то более точный перевод будет «поить кого-то чем-то». Предложенные переводные эквиваленты являются непосредственными потомками СК, в котором лежит лексический класс «давать», и наследуют его модель, поэтому выделение более чем одного значения глагола 喂 wèi представляется нецелесообразным. В большинстве контекстов подходит перевод «кормить», поэтому глагол 喂 wèi был размещен лексикографом в СК ‘TO_FEED’, а случаи отладки перевода были прокомментированы.

Таким образом, по результатам обработки первых трех приоритетов словаря HSK8000 и заведения их в СИ в ходе перегруппировки, добавления и исключения значений на 01 апреля 2012 было описано 7852 лексических значения из 4590 вокабул.

2.3. Типологические особенности синтаксического выражения глубинной модели

Работа над словарной статьей в проектной среде включала следующие задачи: верификация, сортировка и ранжирование исходных словарных значений из словаря HSK8000, составление подробного лексикографического комментария к каждому значению согласно его типу синтаксической парадигмы, приписывание смыслоразличительных в рамках СК семантем и атрибутов грамматических свойств.

В мере обработки языкового материала лексикографический комментарий становился более формализованным, особенно в части описания диатезных схем и залогов. Поверхностная реализация глубиной модели стала описываться в терминах поверхностных позиций, выделенных исходя из представлений об оптимальном синтаксическом разборе. Грамматические свойства лексики маркировались граммемами. По мере обработки все большего объема лексики расширялась номенклатура грамматических категорий и поверхностных позиций. На данный момент для китайского языка определены 72 поверхностные позиции и 30 грамматических категорий (ГК). Используемые на данный момент ГК описывают следующие свойства: тип синтаксической парадигмы, тип объекта при предикате, модель редупликации, возможность залоговых трансформаций, тип имени и счетного слова, возможность слова выступать в той или иной синтаксической роли, линейный порядок в предложении, особые случаи подключения зависимых, модель деривации.

Типологическими особенностями китайского языка является относительно жесткий порядок слов и ограниченные возможности подключения зависимых в линейном порядке относительно ядра. В китайском языке при глаголе в препозиции и постпозиции к ядру, как правило, одновременно может выражаться не более двух актантных ГП. По этой причине семантическая модель китайского значения и его переводного эквивалента в европейском языке зачастую отличаются несимметричностью. При переводе предложения с многоактантной структурой на китайский язык приходиться прибегать к следующим переводческим трансформациям:

- редупликации глагола;
- выносу дочерней зависимой в прадочернюю;
- введению сочиненной или зависимой предикации;
- выносу зависимой в топик.

Пример 1

Русский глагол *«ударить»* допускает одновременное выражение при ядре четырех ГП: [Agent] (кто ударяет), [Object] (кого ударяет), [PlaceOfContact] (куда ударяет) и [Instrument] (чем ударяет).

(1) [Agent: Сяо Ли] ударил [Object: Сяо Вана] [Instrument: кулаком] [PlaceOfContact: по голове].

Китайский переводной эквивалент глагола 打 (dǎ) допускает только одну зависимую в постпозиции и две в препозиции. При необходимости описать аналогичную ситуацию в китайском языке нужно либо удвоить глагол, либо вынести зависимую [PlaceOfContact] в прадочернюю к [Object].

(2) [xiǎo lǐ 小李]/[yī quán 一拳]/[dǎ 打]/[xiǎo wáng 小王]/, /[dǎ 打]/[zài 在]/[tóu 头]/[shàng 上]

Сяо Ли/один/кулак/ударить/Сяо Ван/,/ударить/предлог/голова/

(3) xiǎolǐ yī quán dǎ zài xiǎowáng tóu shàng
小李//一/拳//打/[PlaceOfContact:/在[/Object:/小王/]/头/上]

Сяо Ли/один/кулак/ударить/предлог/Сяо Ван/голова/послелог

Пример 2

В китайском языке ГП [Instrument] при глаголах движения не выражается непосредственно при ядре. Выражение этой зависимой возможно только посредством дополнительной предикации. При этом в входном и выходном языке мы будем иметь две разные глубинные структуры:

(4) [Agent: Он] вернулся [Locative_FinalPoint:домой] [Instrument: на самолете]

(5) [Agent: tā 他/[MannerOfPositionAndMotion:zuò 坐/[Locative:fēi 飞jī 机]]/huí 回/[Locative_FinalPoint:jiā 家]

Он/сидеть/самолет/возвращаться/домой

Таким образом, в переводе мы получаем — «он, сидя в самолете, вернулся домой».

Пример 3

Часто при невозможности выразить зависимую при ядре, она выносится в топик.

(6) [Sphere:shuōhuà 说话/]tā 他/dōu 都/hěn 很/zǐxì 仔细

разговаривать/он/всегда/очень/внимательный

Он очень внимателен [Sphere: в том, что говорит].

Пример 4

При глаголе bō 播 «сеять» выражение ГП [Locative_IntialPoint] и [Locative_FinalPoint] невозможно, однако, оно допустимо при его двусложном синониме bōzhǒng 播种 (сеять).

(7) cóng cūn zǐ nán biān dào běi biān dōu bōzhǒng le xiǎomài
[从/村子/南/边/][到/北/边]都/播种/了/[小麦]

с/деревня/юг/край/до/север/край/все/PFV/пшеница

[С южного края деревни] [до северного края] все засеяно пшеницей.

Когда проблема несимметричности носит универсальный и регулярный характер, преобразования исходной структуры не комментируются, но фиксируются как актуальные проблемы семантических и синтаксических описаний китайского языка, которые будет решаться на дальнейших этапах работ. На данный момент лексикограф только указывает случаи несимметричности для основных (актантных и наиболее частотных) глубинных позиций для лк и СК, поясняет, с каким синонимом из СК она может

быть выражена, а также предлагает возможные варианты переводческих трансформаций.

Заключение: перспективы разработок и использования словаря

На данный момент в СИ описаны порядка 8000 тысяч значений и обработана лексика первых трех приоритетов словаря HSK8000, что, по мнению составителей, покрывает 91% лексики современного китайского языка. Такой объем работ позволяет считать разработку лексико-семантического словаря ядерной лексики китайского состоявшейся.

В ходе 4-летних работ был создан уникальный продукт, не имеющих аналогов в мире, — это первый системный переводной лексико-семантический словарь ядерной лексики китайского языка. В словаре представлена подробная информация по семантической и синтаксической сочетаемости, словоупотреблению, идиоматике. Подход к описанию менялся и развивался в ходе разработок, поэтому часть комментариев имеет смысл расширить с учетом приобретенного в ходе работ опыта. Кроме того, формат и содержание словарной статьи может быть пересмотрен, сокращен или расширен с учетом возможного применения этого продукта. На данный момент перспективными представляются следующие сценарии использования словаря:

1. Самостоятельный продукт — переводной словарь грамматической сочетаемости

В зависимости от аудитории (изучающих китайский язык или профессиональных лингвистов), примеры в словаре могут быть снабжены транскрипцией, проглоссированы, переведены на любой иностранный язык, а семантическая и синтаксическая модель упрощена или описана более подробно и формализовано.

2. Лексическая база для автоматической обработки китайского языка

Данный словарь изначально разрабатывался под задачи системы анализа текста, создаваемой в компании ABBYY. Специфика китайского языка и недостаточная проработанность грамматической теории стала причиной того, что работа над лексикографическими описаниями стоилась по принципу поиска наиболее оптимального синтаксического анализа и учёта всех грамматических явлений, что отражено в проектной методике. Данный словарь может быть использован как лексическая база для задач автоматической обработки китайского языка, разрабатываемых не только в компании ABBYY, но и за ее пределами.

3. Техническое задание для учебной и/или научной грамматики.

В результате работы над словарем и обработки большого лексического материала стали очевидны несовершенство и непоследовательность имеющихся на данный момент учебных и научных грамматик китайского языка. Кроме того, для решения задач автоматической обработки текста нужна специальная грамматика, словарь и база данных грамматических свойств. Пример

грамматического словаря такого типа — словарь грамматической информации, созданный в Институте Компьютерной Лингвистики Пекинского университета под руководством профессора Юй Шивеня [16].

Разработка лексико-семантического словаря ядерной китайского языка заложила крепкую основу для будущих системных описаний грамматического строя китайского языка, что могло бы лечь в основу принципиально новой грамматики языка, ориентированной на решение разнообразных прикладных, педагогических и научных задач.

References

1. *Anisimovich K. V., Druzhkin K. Y., Minlos P. R., Petrova M. A., Selegey V. P., Zuev K. A.* (2012). Syntactic and Semantic Parser based on ABBYY Compreno Linguistic Technologies. Kompyuternaia Lingvistika I Intellektual'nye Technologii: Trudi Mezhdunarodnoj Konferentsii "Dialog 2012" T. 2 [Computational Linguis- tics and Itellectual Technologies: Proceedings of the International Conference "Dialog 2012" Vol. 2]. Bekasovo, pp. 91–103.
2. *Apres'an Yu. D.* (1995). Integral'noe opisanie yazika i sistemnaya leksikografiya [Integral description of language and systemic lexicography]. Moscow.
3. *Chu-Ren Huang, Ru-Yng Chang, Hsiang-Pin Lee.* Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO, Academia Sinica, available at: http://bow.sinica.edu.tw/file/BOW040528WN01.pdf
4. *Claudia Ross, Jing-heng Sheng Ma* (2006). Modern Chinese Grammar. Routledge Modern Grammar Series, London and New York.
5. *Dragunov A. A.* (1952). Issledovaniya po grammatike sovremennogo kitajskogo yazika [Studies on Modern Chinese Language Grammar], Moscow and Leningrad.
6. *Kam-Fai Wong, Wenji Li, Ruifeng Xu, Zheng-sheng Zhang* (2009). Introduction to Chinese Natural Language Processing, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers.
7. *Liu Yang, Yu Shiwen, Yu Jiangsheng* (2005). CCD Yuyi zhishikude gouzao yanjiu [Study on the Construction of CCD]. Xiaoxing weixing jisuanji xitong [Mini-micro systems], Vol. 26, Ed. 8.
8. *Van Li* (1989) Parts of speech [Chasti rechi], Novoe v zarubezhnoj lingvistike [New Trends in Foreign Linguistics], Moscow, Vol. 22., pp. 37–53.
9. *Yip Po-Ching, and Don Rimmington* (2004). Chinese: A Comprehensive Grammar. Routledge Comprehensive Grammars Series, Routlenge.
10. *Yu Jiangsheng, Yu Shiwen* (2002). Zhongwen gainian cidiande jiegou. [The Structure of Chinese Concept Dictionary]. Zhongwen xinxi xuebao [Journal of Chinese Information Processing, 2002, Vol. 16, Ed.4.
11. *Zhendong Dong, Qiang Dong.* HowNet, available at: http://www.keenage.com/

Dictionaries

1. 汉语8000词词典. Hanyu 8000 ci cidian [A Dictionary of Chinese Usage: 8000 words], Beijing, 2007
2. 汉语动词用法词典. Hanyu dongci yongfa cidian [A Dictionary of Chinese verbs usage], Beijing, 2005
3. 汉语形容词用法词典. Hanyu xingrongci yongfa cidian [A Dictionary of Chinese adjectives usage], Beijing, 2003
4. HSK词语用法详解. HSK ciyu yongfa xiangjie. [A Guide to the Usage of HSK Vocabulary], Beijing, 2006
5. 现代汉语语法信息词典详. Xiangdai hanyu yufa xinxicidian xiangjie [The Grammatical Knowledge-base of Contemporary Chinese — A Complete Specification], Beijing, 2002.

ШОЛОХОВ ИЛИ ВСЕ-ТАКИ КРЮКОВ? НЕФОРМАЛЬНЫЕ ПРОЦЕДУРЫ ПРИ УСТАНОВЛЕНИИ АВТОРСТВА «ТИХОГО ДОНА»

Михеев М. Ю. (m-miheev@rambler.ru)

НИВЦ МГУ, Москва

В статье на конкретных примерах методом сравнения контекстов разбираются совпадающие ключевые слова, отдельные выражения и целые фрагменты из известных на сегодня текстов Крюкова и Шолохова. Для каждой выявленной совпадающей пары подбирается формальный образец, в который вписываются только структуры контекстов Шолохова и Крюкова и никакие другие. При этом поверка уникальности совпадений осуществляется поиском по Национальному Корпусу русского языка или полным текстам в Гугле. Отбираются только те, которые с наибольшей точностью повторяют друг друга, не встречаясь у других авторов. Перечислены виды совпадений — единичные («точечные»), множественные (состоящие из множества конструкций), а также «комбинаторно» распределенные — по разным местам в текстах одного и другого автора. В каждом случае оценивается вероятность того, что первоначальным автором выражения мог быть Крюков.[1]

Ключевые слова: авторство «Тихого Дона», Шолохов, Крюков, ключевые слова, уникальное словосочетание, сравнение контекстов, Национальный корпус русского языка, виды совпадений

Проблема авторства «Тихого Дона» (и сомнения в авторстве Шолохова) известна давно. Моя точка зрения на этот вопрос: плагиат исключен просто потому, что бóльшая часть текста романа явно переписана «стилем» Шолохова, хотя в начале существенное использование чужого текста (текста Федора Крюкова) несомненно было. — Но даже если значительная доля и языковых красот романа, и отдельных поворотов сюжета, а возможно самого замысла произведения изначально принадлежала Крюкову (далее сокращенно: К) — все равно текст последнего почти безвозвратно погребен под мощнейшей струей — нового языка и стиля, т. е. языка и стиля нового времени и нового автора(ов). Дело даже не в том, что «авторов», прилагавших руку для приведения романа к виду написанного «красным Львом Толстым»[2] было много (мои наблюдения как раз

---

[1] Доклад публикуется в сокращенном варианте. Полная версия доступна на сайте конференции «Диалог».

[2] Это выражение М. Зощенко мне кажется вполне применимым к поставленной Ш-м (сознательно или бессознательно) задаче — переделке «Тихого Дона» в роман соцреализма.

показывают, что стилистически «проект-писатель-Шолохов» достаточно един и довольно-таки последователен[3]): но все же тут не один только **ухаб**, по выражению Солженицына[4], «между вторым и третьим томом» (под «ухабом» имелся в виду разрыв, с одной стороны, между частями с 1-й по 5-ю, вышедшими в двух книгах в 1928 году, и частью 6-й, т. е. 3-м томом, с другой). Речь, конечно, может идти об «отслоении текста», задачу которого закономерно поставила литературовед Д* (И. Н. Медведева-Томашевская: она, кстати, усмотрела сам «ухаб» в романе несколько ранее, чем обозначил его Солженицын, — уже после 28-й главы в 5-й части). Но такое «отслоение» — задача чрезвычайной сложности и с ней пока никто не справился[5].

Вспомогательные понятия

1. лексическая **Конструкция** — слово или сочетание нескольких слов в конкретных формах, задаваемых языком, в которой слова могут быть приведены к «нормальному», т. е. словарному, виду, по сравнению с конкретными текстами; 2. комбинаторный **Образец** — некоторое множество лексических конструкций, связанное определенными отношениями: во-первых, — это конъюнкция, *объединение*, в значении: ‘и то, и другое, A и B, в любом порядке следования с возможностью постановки любого C между ними’ (далее обозначается + , плюсом), во-вторых, — *произведение* (обозначается как · или вообще без какого бы то ни было знака, как для умножения в алгебре), в значении: ‘то и другое, A и B, но в данном закрепленном линейном порядке этих элементов, с возможностью вставки между ними любого C’; знак / в образце — для разделения синонимов, или вариантов ключевых выражений, которые встречаются у одного из авторов в разных текстах, при самоповторах; знак // в образце отделяет выражения одного автора от другого; знак = в иллюстрациях отделяет примеры одного автора от другого; 3. **неполный Вариант** комбинаторного образца появляется тогда, когда в тексте встречается видоизмененная конструкция, близкая по составу к исходному образцу, но с какой-то модификацией (например, состоящей в сокращении числа элементов в образце или изменении отношений между ними); 4. проверка **Уникальности** образца достигается проверкой встречающихся у того или иного автора вариантов — на совпадение по всему множеству доступных текстов либо в Национальном корпусе русского языка (НК), либо, когда этого ресурса недостаточно, поиском по всем русскоязычным текстам в Гугле: если ни один другой вариант, кроме вариантов Ш-а и К-а, не подводим под данный, то образец признается уникальным

---

[3] Здесь я никак не могу согласиться со многими явными утверждениями и неявными презумпциями Бар-Селлы (2005, 2008).

[4] Из письма А. И. Солженицына — И. Н. Медведевой-Томашевской 1971 (Медведева 1993).

[5] Одна из интересных, но так и не завершенных попыток в (Смирнов 2001).

и может рассматриваться в качестве аргумента в пользу происхождения обоих текстов из единого источника, т. е. авторства обоих текстов, отдаваемого К-у.

А. «Точечные» совпадения: метафоры, сравнения, эпитеты

Наиболее частые совпадения, если оценивать статистически, — в диалектных словах и местных донских речениях; это неудивительно, принимая во внимание знакомство обоих с диалектом, так что здесь об этом говорить не будем[6]. Наиболее простое из нетривиальных совпадений в текстах К-а и Ш-а — это совпадение в структурах единичных метафор, сравнений или даже простых эпитетов.

А1. например, такое соответствие — одного места из текста К-а (в очерке 1914 г.) одному же месту у Ш-а, в данном случае только в ТД, — с повтором в обоих единой структуры, или, в моих терминах, единого образца (выделяю его элементы жирным) 1=1:

(облака // солнце) + (словно // как) с похмелья (1914 = ТД 2:21)[7]. Вот примеры (снова вначале К, потом Ш): *Серые неуклюжие облака тяжело клубились за крышами домов, над городским садом и за рекой. Грузно ворочались, лезли друг на друга, ползли вширь, рыхлые, мутные, словно с похмелья, бестолково толпились….* («Мельком» 1914) = (ТД 2:21) *Ветер нес по площади запах конской мочи и подтаявшего снега. Невеселое, как с похмелья, посматривало солнце. # Группа офицеров задержалась около казака …* — в 1-м черновике ТД (Рукопись, Часть 2, с.92), в романе еще не было обсуждаемой фразы про похмелье, она отсутствовала. Были только предшествующая и следующая за ней, а в дальнейшем кроме добавления самой фразы всё выглядит так, как потом и в печатном варианте. То есть фраза с интересующим нас сравнением каким-то образом была вставлена при редактировании — скорее всего, самим Ш-м (хотя возможно будучи заимствована, или скопирована с К-го черновика? — ср. с похожей на нее фразой в очерке К-а «Мельком» 1914, приведенной выше). Конечно, полным совпадением эти фразы не являются, но в них несомненно совпадают образы сравнения и, частично, сами объекты, а основания сравнения при этом разные: в первом случае, это *облака* мутные, неуклюжие, грузные, налезающие друг на друга — *словно с похмелья*, а во втором уже *солнце* невесело посматривающее — *как с похмелья*.[8]

---

[6] О них — в статье, отданной в печать (Михеев 2012).

[7] Такое соответствие, когда с одним из контекстов среди всех известных на сегодня текстов К-а сопоставим ровно один контекст у Ш-а, буду обозначать формулой 1=1 (цифры в такой формуле будут обозначать тексты, по их хронологии, слева — К-ие, справа — Ш-ие).

[8] Благодарю Л. Л. Федорову за совместные обсуждения многих конкретных примеров этой статьи и ценные критические замечания.

Так разве можно вторую метафору считать «списанной» с первой? — О буквальном «списывании» речь не идет. Но возможно это авторский самоповтор: если «предтекст», или как его иначе называют, «протограф», положенный в основание ТД, был написан действительно К-м, то повтор некоторого числа собственных метафор, сравнений, типовых сюжетных связок — вещь вполне естественная (это именно то, что получило у Мезенцева название — «текстологический конвой»[9]). О сходстве говорит структура сравнений, используемых К-м и Ш-м (договоримся все-таки считать, что автором ТД, пока не доказано обратное, являлся Ш). Выше эта структура и записана формально в образце (А1). В нем слова *как* и *словно* — практически одно и то же. Несколько сложнее отождествить объекты сравнения, т. е. *облака* и *солнце*, но все же и это возможно, на основании метонимической смежности — и то, и другое принадлежат к «небесным явлениям».

Ну, а вот аргументы против идеи заимствования и преемственности таких сравнений (т. е. против допущения, что они вышли из-под пера одного автора): во-первых, сама фраза в черновике отсутствовала и была вписана уже только в беловик — очевидно, в результате какой-то скрытой от нас правки рукописи. Но: кажется естественным считать, что более вероятно принадлежали перу К-а те фрагменты, которые внесены в рукопись изначально или все-таки прочитываются под слоем правки, а исправления, которые Ш вносит в текст (или кто-то за него?) — будь то в черновик, беловик, или в текст на каком-то промежуточном, скрытом от нас, читателей, этапе (скажем, еще в голове его создателя, так сказать, между черновиком и беловиком), — естественнее отнести, по-моему, к — авторски Ш-им. Ну, или же приходится предполагать, что при «переписывании» с чужого текста в каких-то местах Ш делал купюры — возможно, просто не будучи в состоянии прочесть оригинал?[10] — а потом, через некоторое время, когда Ш еще раз возвратился к чужому протографу, он смог восстановить оттуда тот фрагмент, вариант которого в его собственном написании как в эскизном «прогоне текста» был им оценен, при сравнении с К-им текстом, как более слабый? Но такое, слишком уж «хитрое» текстопорождение, в постоянном как бы «соцсоревновании» с автором-предшественником, на мой взгляд, все-таки маловероятно.)

Таким образом, хоть и с некоторым сомнением, но совпадение по образцу *(облако/солнце) + как с похмелья* можно признать уникальным и отнести к аргументам в пользу того, что автор обоих текстов — изначально один и тот же. На самом деле, таких примеров много. Приведу еще два.

А2. **всадники (как // похожие на) + нахохленные** (**воробьи // птицы)** (1913 = **ТД** 4:20). Здесь опять-таки совпадение выражений, укладывающихся в единый образец «один к одному» — как и ранее, обозначу это формулой: 1=1.

---

[9] М. Т. Мезенцев должен был быть упомянут уже во введении в проблему. Этот его термин, в свою очередь, заимствован — с некоторым переосмыслением — у Д. С. Лихачева (Мезенцев 1998).

[10] Хотя у К-а в самом деле непростой почерк.

В восемь утра, в сером полусвете, сквозь тонкую, зыбкую сетку дождя я вижу: верхом на мокрых тонконогих лошадках с острыми спинами подъезжают к станичной школе оригинальные всадники *— маленькие, намокшие, нахохленные, как озябшие воробьи* («В глубине» 1913) = (**ТД** 4:20): *Всадники, похожие на нахохленных черных птиц, ехали, надвинув высокие папахи, зябко горбились в седлах, кутали в башлыки маслено-смуглые лица*. — Как же так? — должен сказать здесь внимательный читатель, знакомый с текстами обоих авторов, — ведь тут описаны совершенно разные события: у К-а это событие вполне мирное и даже бытовое, приезд учеников в школу, а в ТД — тайное бегство из Быховской тюрьмы генералов во главе с Корниловым. Помимо всего прочего, само сравнение человека с *нахохленной* птицей просто тривиально. И тем не менее то, что в обоих случаях объектом сходного сравнения оказались именно *всадники*, т. е. совпадение набора конструкций, как бы подобранных по единому образцу, совершенно уникально: ни у кого во всем НК такое не повторяется; нет совпадений и при поиске в Гугле. Это позволяет предположить, что у приведенных фраз автор один. Конечно, самих по себе сочетаний, построенных по образцу *нахохленный + (воробей / птица)* — просто пруд пруди, но вот *всадников*, которые бы сравнивались с ними, почему-то больше нет ни одного. И этот пример тоже — на чашу весов авторства протографа ТД за К-ым.

Несколько более экзотична метафора, заключенная в составе следующего микросюжета:

А3. **(глядеть // оглядывать) + из-под + (белоснежный // белый) лопух шляпы** (1916 = **ТД** 3:8) — и здесь формула снова: 1=1. Совпадает не только метафора, но одновременно еще и несколько сюжетных деталей[11]. В обоих текстах действие происходит во время войны. Только если в рассказе К-а речь шла о монастыре в Галиции, где монахинями организован госпиталь (это на территории бывшей Австро-Венгрии; женщины говорят между собой очевидно по-польски): *У настоятельницы на голове белоснежный накрахмаленный лопух величиной с дамский зонтик. Под этим лопухом сизое, озябшее лицо старушки, с носом вроде созревающего баклажана и толстыми, строгими губами, очень смахивает на великолепный гриб-дождевик изукрашенный ветрами и солнцем. Из-под другого лопуха, поменьше, лукаво глядят карие веселые глазки маленькой, изящной шаритки* [и там же, через 3 страницы:] *Белые лопухи на головах монахинь затряслись, закачались: # — До видзенья, пани... до видзенья!..* («Группа Б» 1916) = у Ш-а в романе русские солдаты-казаки косят траву на поле хозяина-поляка: (**ТД** 3:8) *Хозяин, под белым лопухом войлочной шляпы, повел их к своей деляне.* [и через ½ страницы — продолжение темы, заданной образцом К-а:] *Иванков затягивал в фуражирку ворох травья. Рядом с ним, посасывая трубочку, стоял колченогий хозяин-поляк. Он сунул руки за пояс, из-под полей шляпы, насупясь, оглядывал косившего Щеголькова*.

[11] Благодарю за указание на него, как и за обсуждение некоторых других примеров С. Л. Рожкова.

Метафора *лопух шляпы* при проверке по **НК** больше ни у кого из авторов, представленных в корпусе за интересующий нас промежуток времени не повторяется. А при этом в **ТД** больше нет ни одной метафоры или хотя бы сравнения со словом *лопух*, зато у К-а они встречались и ранее, и после:

У Ш-а нет других метафор с этим словом не только в **ТД**, но и в **ДР**, единственное с ним сравнение — только в **ПЦ**, когда Разметнов укоряет Давыдова: *… все разговоры да разговоры «кулака унистожить», а он растет из года в год, как лопух, и солнце нам застит* (**ПЦ** 1:4). Но это, согласимся, сравнение совсем другого рода, нежели *лопух шляпы*: не уподобление цветку по форме, а здесь возникают уже совсем другие коннотации — неотвязность, невозможность избавиться от сорной травы…

То есть здесь, в этой, признаем, довольно «избитой» литературной метафоре, возможно и ее независимое порождение обоими авторами, но все-таки в составе целостного образца (*глядеть из-под* +) никем другим она не использовалась. Так что и этот пример — аргумент в пользу возможности единого авторства.

Б. Сложно-составные совпадения

В последнем из разобранных выше примеров мы уже столкнулись с тем, что Образец может воспроизводиться самим автором в виде неполного Варианта — когда *белые лопухи* в тексте К-а были повторены в том же тексте через 3 страницы, но — в неполном виде (или у Ш-а, не только в ТД, но и в ПЦ). Несколько более сложное дело — но, как представляется, и более надежное, так как более доказательное — нахождение совпадений сразу по нескольким описываемым предметам, действиям или ситуациям. При этом сами конструкции могут описываться как с помощью метафор, так и быть сугубо «денотатными», т. е. бытовыми, нехудожественными, — когда уникальность Образцу обеспечивает неповторимость одних и тех же его элементов.

Б1. **(молодежь // парни) + (заигрывать с // щупать)** + **девки** + **в (церковь + ограда) // в церковной ограде** (1912 = ТД 2:16): 1=1. Тут перед нами описание ситуации без метафор и без всяких поэтических сравнений — некий простенький сюжетный эпизод: в текстах **К**-а он встретился один единственный раз — в рассказе священника с жалобой о том, как он взывает к совести своих прихожан, читая им проповедь против пьянства, а они, мол, особенно молодые, никак этим призывам не внемлют: *… произносил проповеди в церкви <…>*, а *молодежь в это время толчется где-нибудь в ограде, с девками заигрывает*… («Без огня» 1912). = Так же единственный раз встречается этот Образец и у Ш-а, только в ТД и — переходя в авторскую речь (следуя в абзаце за описанием погоды): (2:16) *В церковной ограде, испещренной блестками талых лужиц, гуртовались парни. Из церкви через распахнутые двери на паперть, с паперти в ограду сползали гулкие звуки чтения, в решетчатых окнах праздничный и отрадный переливался свет, а в ограде парни щупали тихонько повизгивавших девок, целовались, вполголоса рассказывали похабные истории.*

В. Множественно-комбинаторные совпадения

(Такие случаи, когда имеет место «вязкий» повтор одного и того же образца — например, в разных текстах К-а, а воспроизводятся они у Ш-а только уже все вместе, в одном ТД.)

В1. (**Шоколадный** цвет // **шоколад) + (пятна / жижа / жидкая грязь / пюре // глина)**[12]. — Здесь имеется четырехкратное упоминание одной и той же конструкции с метафорой в разных текстах К-а (1912–1916), а им соответствует однократное воспроизведение в **ТД** (только в 3-й части). Для К-а это одна из постоянно повторяющихся метафор: *грязь* для него похожа на *шоколад*, — но у **Ш**-а, в **ТД**, они же, два этих реальных предмета, существующие все-таки отдельно, *шоколад* и *глина*, хотя последняя сравнивается в конце концов с шоколадом, но только по цвету: (1912, 1914, 1915, 1916 = **ТД** 3:17; 4:3). В текстах **К**-а этот образец — явный маркер, который в общей сложности возникает не менее 7 раз («Офицерша» 1912, «Мельком» 1914, «В Азербайджане» 1915 — дважды, «Группа Б» 1916 — а в последнем тексте даже трижды).

В первый раз — у К-а в «Офицерше», где с шоколадом сравнивается глина, которой перепачкано лицо и одежда подновляющей свой курень казачки (тремя абзацами ранее этого места персонаж-повествователь наблюдал, как героиня штукатурит, обмазывая глиной стену): 1. *он с минуту смотрел на нее, ширококостную, рябую, до бровей забрызганную шоколадными пятнами* («Офицерша» 1912). Еще раз повторено то же сравнение и в другом очерке, через два года, где описано уже путешествие в лодке по реке: 2. *Мы причалили к зеленой муравке. Я прыгнул и… увяз — попал в болотину, брюки до колен покрылись слоем жидкой грязи шоколадного цвета. Огорчился и сконфузился… Мужичок успокоительно говорил: — Тут-таки вязко. Да ничего, обсохнет. Как говорится: не сало — помял и отстало…* («Мельком» 1914). В следующем году — уже с небольшой лексической модификацией метафоры, на вместе *грязи* как денотата оказывается еще одно ее сравнение, или метафора, — *пюре*, хотя и сам денотат потом тоже будет назван, для определенности; кроме *пюре* к эпитетам грязи добавлен здесь еще и — *глянец* (позже он тоже будет повторен, в еще одном тексте): 3. *А дорога была — сплошное шоколадное пюре, местами жидкое, отливающее радужным глянцем, местами — покруче, темное и матовое. Фаэтон, плавно покачиваясь, медленно плыл по этому глянцевому мес<ив>у … И постепенно в душе чувство содрогания и изумления перед неисчерпаемым обилием и царством грязи сменялось чувством успокоительного примирения: что ж, и в грязи люди живут…* («В Азербайджане» 1915). Ну, а последний раз — уже троекратное обращение к тому же образу в одном и том же тексте, на протяжении полутора десятков страниц: 4. *военные фуры, двуколки — все забрызганное глянцевой шоколадной грязью*; тут еще повторяется и причастие *забрызганный*; 5. *Шипя шоколадною грязью, жужжа, брызгаясь, пронесся автомобиль*; 6. *На шоссе стояла шоколадная жижа в полколена* («Группа Б» 1916). = Всему этому изобилию соответствуют у **Ш**-а только один пример в рамках указанного

[12] Отмечено в словаре параллельных мест А. Ю. Чернова.

образца. При лексически уникальном совпадении сами ситуации опять-таки совсем не совпадающие, в первой: шоколад упомянут реально, как лакомство, который в теплых пальцах отца Мелехова становится расплавленной *жижей*, и — отдельно от этого — хотя и контактно, в следующем же предложении, упоминается глина, о которую герой вытирает руку: (**ТД** 3:17) *Пантелей Прокофьевич давил в горсти очищенную конфету. Шоколадная вязкая жижа ползла у него меж пальцев. Он вытер ладонь о коричневую крошкую глину насыпи, молча стал закуривать.*

Итак, четыре раза повторенный у К-а полноценный образец превращается у Ш-а в один единственный, в ТД, обозначим это формулой: 1+2+3+4=1.

В2. **(Задумавшиеся / после раздумья // в раздумье) +** (галки / куры // **курица).** Перед нами случай, когда образец, состоящий из двух конструкций с уникальной метафорой, при этом повторенный трижды, с теми или иными вариациями, в трех текстах К-а, затем воспроизводится у Ш-а, — тоже только в одном месте, в ТД, т. е. три примера сводятся к одному: предварительная формула этого такова: 1+2+3=1. У **К**-а в 3 текстах, написанных почти подряд — в 1907-м, 1908-м и 1909-м годах, а у **Ш**-а в одном месте — только в **ТД**: 1. … *золотисто-багряные закаты с алыми стенами станичной церковки и задумавшимися галками на крестах* («Обыск» 1907); 2. … *куры с испуганным кудахтаньем метались на плетни и после короткого раздумья ныряли во двор* («Жажда» 1908); 3. (мальчик Дениска видит отраженную в лужице) — *задумавщуюся курицу и черную ветку старой груши* («Зыбь» 1909). = **ТД** (3:2): *На обезлюдевшем дворе ходила пестрая, с подрезанным хвостом курица и, не зная того, что назавтра помышляет повар приготовить из нее суп пану управляющему, походя копалась в навозе и клохтала в раздумье, где бы положить яйцо.*

Очевидно, что для К-а *задумавшаяся* птица — опять-таки некий ключевой элемент «текстологического конвоя», поскольку она трижды возникает в фокусе писательского внимания. Но вот у Ш-а ни один из представителей пернатых, кроме злосчастной (ни о чем *не подозревающей*) курицы, т. е. ни галки, ни так любимые им хищники — коршуны или соколы — не описываются как *задумавшиеся* или же пребывающие *в раздумье*.

Приведенной вначале формуле для контекстов К-а и Ш-а, соответствующих данному Образцу, можно придать более точный вид, указав для определенности (это пригодится в дальнейшем) в нижних индексах при каждом из контекстов — количество элементов в составе образца: $1_2+2_2+3_2=1_2$ (то есть все три варианта уникальны, так же как вариант в ТД, имея в своем составе по два элемента). Здесь мы видим, что подобного рода совпадения отличаются тем, что «насыщение» их «слева» существенно богаче, чем «справа». Казалось бы, почему?

Г. «Комбинаторные» совпадения

В примерах ниже образец, регулярно употребляющийся кем-то одним из авторов, у другого в целостном виде представлен только один раз или же бывает «собран» из нескольких лексических конструкций порознь.

Г1. **(глядеть // таращить глаза на) + обтянутый (рубахой // жилеткой) + круглый живот**[13] (1914 = 1928). Это образец из трех слагающих его конструкций, без настоящих метафор; у К-а он потом повторен еще раз, в тексте более позднем и в варианте неполном. 1. *Максим Семеныч с усилием таращит глаза на обтянутый рубахой круглый живот* («Тишь» 1914) = а у Ш-а он встречается лишь однажды, в **ТД** (2:1). Про сына купца Мохова, Владимира, который идет к отцу жаловаться: *глядя на круглый отцовский живот, обтянутый чесучовой жилеткой, уже решительно продолжал…* Ситуации различаются тем, что в первом случае, у К-а, сам автобиографический герой повести, ветеринар Карпов, — *таращит глаза* — на собственный живот. При этом позднее, уже в военном рассказе «Группа Б» (1916) встречаем у К-а и неполный вариант, без конструкции *глядеть / таращить глаза на* — там упоминается некий доктор, который: 2. *стоял, выпятив круглый живот, обтянутый кожаной курткой, курил папиросу за папиросой* (но даже этот вариант также оказывается уникален, по НК).

Таким образом, оба подходящие под описанный Образец примера — К-ий и Ш-ий — в полной форме, т. е. с тремя составляющими конструкциями, уникальны, и ближе всего к ним — еще один К-й, уже неполный, только с двумя конструкциями (он не уникален, по Гуглу). Но вот в текстах Ш-а нигде в другом месте подобного не встречается. Здесь формула соответствия такова (неполный вариант в скобках): $1_3+(2_2) = 1_3$.

Г2. В следующем случае имеем соответствие множества самоповторов у К-а — одному единственному месту в ТД[14] — с еще одним, неполным, уже из **ПЦ**: **(медовый // медвяный)** + (**запах / пахнет // пахнуло)** + (**цветы // цветущая**) **+ тыква**[15]. Выстроив тексты К-а и Ш-а по годам: 1909, 1910, 1913, 1916, 1918, 1919 = **ТД** 6:61 (1932)[16]; **ПЦ** 2:11 (1960). А вот формула самого образца, забегая несколько вперед, такова: $1_4+(2_2)+(3_3)+(4_2)+(5_2)+(6_2)=1_4+(2_2$). С одной стороны, этот уникальный образец у К-а повторяется в шести текстах — у Ш-а только в двух местах, в ТД и ПЦ, второй из которых неполон. Но у К-а среди них также все кроме первого неполные, и даже наиболее «насыщенный» из всех все равно «уступает» Ш-у. Вот соответствующие контексты К-а: 1. *Густой медовый запах шел от крупных золотых цветов тыквы с соседнего огорода* («Зыбь» 1909) — здесь четыре совпадения с образцом (они снова помечены нижними индексами); 2. *уютные стены, которые так знакомо пахнут печеной тыквой и дымком* («Мать» 1910) — всего лишь два совпадения; 3. *Бессильно свернулись золотые, медовые цветы тыквы* («Отец Нелид» 1913) — три

---

[13] Было отмечено у А. Чернова и Д. Доброва.

[14] Этот пример был отмечен уже многими исследователями Крюкова — А. Ю. Черновым, Дм.Добровым и др.

[15] *Медовый* и *медвяный* считаю возможным в данном случае отождествить — применительно к запаху тыквы или ее цветов.

[16] 61-я глава 6-й части ТД была опубликована в 1932 г.

совпадения; 4. *Пахнет кизячным дымком и печеной тыквой* («Ползком» 1916) — тут опять только два; 5. *хуторские улички с кучами золы в центре, с запахом кизяка и печеной тыквы* («Свежо предание 1918) — также два; 6. *запах кизека и тыквы борется с чудесным запахом осеннего листопада* («Цветок-татарник» 1919) — и здесь тоже два.

А вот пример Ш-а — единственный в **ТД**, наиболее «насыщенный», вбирающий в себя все содержание, все конструкции этого образца, даже с избытком: (6:61) *с огородов пахнуло медвяным запахом цветущей тыквы*. — Здесь формально даже больше, чем в наиболее насыщенном контексте К-а, т. е. 5 совпадений! (но *запах* и *пахнуть* — синонимы).

Д. Распределенные по разным местам совпадения

Наконец, встречаются такие случаи, когда образец собирается из нескольких конструкций, или выражений порознь — когда и у того, и у другого автора имеют место «вязкие» повторы одной и той же конструкции, рассеянные по разным текстам или же по разным местам в одном тексте:

Д1. **Веер** + (**раскинуться // расходиться) + (багряный // багрянец) + (пыльный** / колонны **пыли** / столбы **пыли // пылиться) +** (облака // тучи) **+ (заря // лучи) + (**на **западе // закатное** солнце**) +** (**мазки / краски // окраска / отсветы / спектр красок**)[17]. («Мать» 1910 = **ТД** 3:14) — здесь точное соответствие (1=1), снова по максимуму вместе 7 из 8 элементов Образца встречаются только у **К**-а в тексте «Мать» (1910 г.) и у Ш-а в **ТД** 3:14 (текст опубликован в 1928-м). Дам их перечисление наряду с неполными вариантами: 1. *Когда последние, красноватые лучи его* (солнца) *погасли на курившихся облаках пыли, стало вдруг очень свежо, даже холодно, захотелось деревни, ночлега* («К источнику…» 1904); 2. *Косые лучи выглянувшего из облаков осеннего солнца пронизывали оригинальные колонны пыли* («Новые дни» 1907); 3. *В прорехи старой крыши сарая лезли горячие лучи солнца. Светлые, чистые колонны пыли, разрытой курами, стояли косыми рядами, наклонившись в сторону улицы* («Зыбь» 1909); 4. *На позеленевшем, низком небе застыла золотисто-розовая пыль зари* («Товарищи» 1909 — из 1-го абз.); 5. (депутат от станицы идет по полю вместе с местным батюшкой) *пошел, окруженный плотным, непроницаемым живым полукругом, от которого пыхало жаром. Песочная мелкая пыль тающим белым столбом плыла над ними* («Шквал» 1909). Через год добавлен мотив «живописных мазков красок»: 6. *Пыль плавала над ярмаркой, и дым, голубой и пахучий, медленно выползавший из харчевен, опоясывал, как шарф кисейный, золотую стену высоких верб и тополей. Низко стояло солнце. Лучи, перерезанные длинными тенями, бросили на площадь косые, схилившиеся колонны, прозрачные и светлые. Тесное шумное торжище было похоже — на волшебный храм: голубой, запыленный свод; зеленые стены с причудливой канвой, расписанной фантастическими мазками лазоревых, пурпуровых, лиловых и золотых красок; дым кадильный; многоголосый хор*

[17] Слова, не выделенные жирным, входят в Образец только у кого-то одного из авторов.

нестройно жужжит, плещется, шумит… ; 7. *Вечер надвинулся. Пыльная заря горит на западе, багряные длинные мазки веером раскинулись на полнеба* («Мать» 1910) = 1. (**ТД** 2:3) *И деды, провожавшие с завалинок пыльный багрянец заката, переговаривались через улицу…*; 2. (**ТД** 3:14) *Желтую щетину жнивья перебирал ветер. На западе корячились, громоздясь, тучи. Вверху фиолетово чернели, чуть ниже утрачивали чудовищную свою окраску и, меняя тона, лили на тусклую ряднину неба нежно-сиреневые дымчатые отсветы; в средине вся эта бесформенная громада, набитая как крыги в ледоход на заторе, рассачивалась, и в пролом неослабно струился апельсинного цвета поток закатных лучей. Он расходился брызжущим веером, преломляясь и пылясь, вонзался отвесно, а ниже пролома неописуемо сплетался в вакханальный спектр красок.*

С привлечением указанных неполных вариантов, схема образца выглядит так: $(1_3)+(2_3)+(3_2)+(4_2)+(5_2)+(6_4)+7_7 =(1_3)+2_7$.

Итак, мы обозрели здесь типы возможных соответствий. Когда будет рассмотрено все их множество по текстам Ш-а и К-а, можно будет точно определить степень зависимости текстов одного от текстов другого. Замечу, во-первых, что подавляющее большинство совпадений оказывается — по тексту ТД, а не с другими произведениями Ш-а, а во-вторых, что последние из рассмотренных нами типов (В, Г, Д) отличаются тем, что повтор образца имеет место или только «слева», со стороны К-а, или преимущественно «слева», а не «справа» (не со стороны Ш-а), «справа» же их значительно меньше. Это, на мой взгляд, свидетельствует о заимствовании фраз, составленных по этим образцам, из текста К-а. Ясно главное: в какой-то степени текст К-а использовался, но все-таки не как объект работы плагиатора. Ш серьезнейшим образом «перепахал» его, постаравшись вмешаться, или «наложить руку», переделать по-своему буквально все литературные находки своего предшественника. Насколько он «улучшил» или «ухудшил» в итоге текст, можно спорить еще долго, это дело вкусов (и идеологических пристрастий) исследователей. Но отделить «влияния» на текст (которые здесь не разбирались: среди безусловно влиявших на автора **ТД** можно указать хотя бы Есенина и Бабеля, при всем несходстве двух поэтик) — однако отделить такие влияния от текста, «использованного» для создании ТД, как оказывается, все-таки возможно.

Reductions

ДР — «Don stories» of Sholokhov [“Donskije rasskazy”]
К — Feodor Dmitriyevich Kryukov (its texts are quoted according to electronic editions: Крюков, Чернов, ФЭБ.)
НК — National corpus of Russian language [Natsionalnyj korpus russkogo jasyka], available at: http://ruscorpora.ru/search-main.html
ПЦ — Sholokhov M. Virgin soil upturned [Podnjataja tselina]
ТД — Sholokhov M. Novel “Quiet Don” [Tikhij Don]], available at: http://ruscorpora.ru/search-main.html

ФЭБ — Fundamental electronic library [Fundamentalnaja elektronnaja biblioteka] http://feb-web.ru/feb/sholokh/default.asp
Ш — Sholokhov M.

References

1. *Bar-Sella Z.* (2005) Literary ditch. Project "Writer Sholokhov" Moscow, 2005. [Literaturnyj kotlovan.]
2. *Bar-Sella Z.* (2008) Notes of the dead man ("Quiet Don": chronology textual criticism). Russian mail. Belgrad 2008, № 1. [Zapiski pokojnika. Russkaja pochta. Beograd, 2008, №1]
3. *Dobrov D.* "Quiet Don" Unfinished novel [Neokonchennyj roman], available at: http://www.dm-dobrov.ru/publicism/don/index.html.
4. *Kryukov F. D.* (2009). Over break. Sketches and articles of the last years of life (1917–1919). Mocow, S. Ptb. 2009 . [Nad obryvom.], available at: http://unipersona.srcc.msu.ru/f-krukov/index.htm
5. *Kryukov F. D.* (2010). Incomplete collected works [Nepolnoje sobranije sochinenij], available at: http://fedor-krjukov.narod.ru/index.htm
6. *Medvedeva I. N.* (1993). "Quiet Don" stirrup. Novel riddles. Moscow, 1993 [Stremja Tikhogo Dona].
7. *Mezentsev M. T.* (1998). Destiny of novels. Samara. [Sud'ba romanov], available at: http://www.philol.msu.ru/~lex/td/?pid=0121&oid=01.
8. *Mikheev. M. Iu.* Compilation or… language clichés? Comparing sets of word collocations characteristic for authors' styles. [Kompil'azija ili jazykovyje klishe] Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010", V. 9 (16), p. 340–349. available at: http://www.dialog-21.ru/digests/dialog2010/materials/html/52.htm
9. *Mikheev. M. Iu.* Couleur local or ornamentalizm in the art prose. Dialect impregnations by Kryukov and Sholokhov [will be published allegedly in 2012]
10. *Pavlovich N. V.* Dictionary of poetic images. 1999 [Slovar' poeticheskih obrazov]
11. *Manuscript 2005. Sholokhov M. Manuscript "Quiet Don". Facsimile edition Moscow-Kiev-Paris, 2005 [Rukopis "Tikhij Don"]*
12. *Smirnov A. A.* (2001) Readings of "Tikhij Don". Readings first, available at: http://www.philol.msu.ru/~lex/td/?pid=0211&oid=02
13. *Chernov A. Iu.* (2011) Feodor Kryukov. "Quiet Don". Materials to the parallel dictionary of dialecticisms [Feodor Kryukov — "Tikhij Don". Materialy k parallelnomu slovariu dialektismov]

ESTONIAN SPEECH SYNTHESIS: APPLICATIONS AND CHALLENGES

Mihkla M. (meelis@eki.ee),
Hein I. (kiisu@eki.ee),
Kalvik M. (mariliis@eki.ee),
Kiissel I. (indrek@eki.ee),
Tamuri K. (kairi.tamuri@eki.ee)
Institute of the Estonian Language, Tallinn, Estonia

Sirts R. (risto.sirts@err.ee)
Estonian Public Broadcasting, Tallinn, Estonia

In the 21st century Estonian speech synthesis has been developed using the more widespread methods and freeware development systems (MBROLA, Festival, eSpeak, HTS). The applications have hitherto been developed mainly in view of the needs of the visually impaired (audio system for reading electronic texts, voicing of subtitles, creation of audiobooks). The major challenges currently facing the Estonian specialists are naturalness of the output speech and expressive speech synthesis. The article is concerned with the issues of statistical modelling of the prosody of synthesized speech and the relations of prosody with other language levels as well as with extralinguistic features. Analysis of the emotion-bound acoustic parameters (pauses, speech rate, formants, intensity and pitch) enable one to model emotions for speech synthesis. In addition, speech synthesis interfaces are discussed. By means of such interfaces users could control the process of speech synthesis, monitor text-to speech transformation, follow text structure and vary the parameters (voice loudness, speech rate, voice pitch) of the synthetic voice in various voice applications.

Key words: speech synthesis, voicing of subtitles, emotional speech, Estonian

1. Introduction

The history of speech synthesis knows various TTS (text-to-speech) applications from reading systems for the blind to modern interactive dialogue systems. Application of the first synthesizers was mainly limited by the machine-like sound and poor intelligibility of the output speech. Moreover, the past half century of history has witnessed several misinterpretations of the essence of speech synthesis and TTS applications. Thinking back to the 1980s with the conferences on automatic recognition of auditory images (автоматическое распознавание слуховых образов — АРСО), it was often complained that the advent of speech synthesizers failed to meet the enthusiastic expectations of the scientific and technological revolution; for many people it rather looked like a fiasco. This was not mainly because the first synthesizers

sounded like robots and were sometimes hard to understand, but most of the disappointment was due to the fact that the synthesizer could not understand human speech. Somehow the sci-fi novels and films had inculcated the idea that a speaking machine should also recognize and understand speech. Doubts were even expressed whether speech synthesis was necessary at all. Today, however, TTS output has become so good that sci-fi film producers have to make a special effort to supply synthetic speech with some additional machine-like characteristics in order to prevent the audience from thinking the robots can understand speech (Taylor 2009). Despite such misinterpretations of the abilities of a speech synthesizer and undue attribution of features, modern TTS systems have quite a lot of applications as well as challenges to face, while speech synthesis contributes, first and foremost, to technologies of speech recognition and understanding (artificial intellect); however, people often find it hard to think of the two speech technological components as separate.

In Estonia such misinterpretation of speech technological abilities and overestimation of speech synthesis has not been a problem. If a language has just a million speakers, its speech technological applications are not too profitable and the problem often is how to find enough users for the applications. The Estonian experience of speech synthesis has a history of over forty years now. Already the advent of the first generation systems was accompanied by active efforts to create user applications, in particular for the visually impaired (Remmel, Tago 1984). About 15 years ago, after a short break, research and development into Estonian speech synthesis was renewed and has been advancing briskly ever since, using various methods. Speech synthesis being one of the vital language technological tools providing for the survival and versatile development of any national language in the 21st century information society, R&D into speech synthesis has enjoyed considerable state support in the recent years.

The article will survey the available Estonian speech synthesizers and their applications (e. g. an audio system for reading electronic texts, the voicing of subtitles and creation of audiobooks). The major challenges currently facing the Estonian specialists are naturalness of the output speech and expressive speech synthesis. The article is concerned with the issues of statistical modelling the prosody of synthesized speech and the relations of prosody with other language levels as well as with extralinguistic features. On the basis of the Estonian Emotional Speech Corpus the acoustic characteristics (pauses, speech rate, formants, intensity, pitch) of some basic emotions (anger, joy, sadness) and of neutral speech are being analysed. The aim is to ascertain the emotional and emotion-specific acoustic parameters enabling one to model emotions for synthetic speech. Other notable applications include smart interfaces (Speech Application Programming Interface), enabling one to control the process of speech synthesis, to monitor text-to speech transformation, to follow text structure and vary the parameters (voice loudness, speech rate, voice pitch) of the synthetic voice in various voice applications.

2. Modules of Estonian speech synthesis

The first Estonian speech synthesizers were parametric, mainly using formant-based synthesis. Over the past decade the research has concentrated on compilation-based

synthesis using donor voices, without, however, completely neglecting the formants. As the local specialists are very few indeed most of our products use ready-made development systems, while he local efforts are mainly focused on language-specific issues.

2.1. Diphone speech synthesizers

Around the turn of the millennium an Estonian concatenative text-to-speech synthesizer was produced (Mihkla, Eek, Meister 1999). The speech units used were diphones drawn from a 1900-strong database of sound-to-sound transitions. Figure 1 represents a block diagram of an Estonian diphone synthesizer, explaining what building blocks or modules are necessary for producing synthetic speech from any text written in Estonian. In principle, the same scheme underlies the TTS conversion with other methods of synthesis as well. Most of the procedures, rules, formalisms, lexicons and databases are language-specific modules, except, at least in a general case, the (double-framed) block of signal processing. The diphone synthesis was performed using the MBROLA engine. Many people with a visual impairment still use the Estonian diphone-based synthesizer. The main shortcomings of MBROLA synthesizers are due to neglect of further development of the project, which, considering the recent developments in hardware is causing compatibility problems (e. g. the ghost echo in the case of multi-kernel processors).

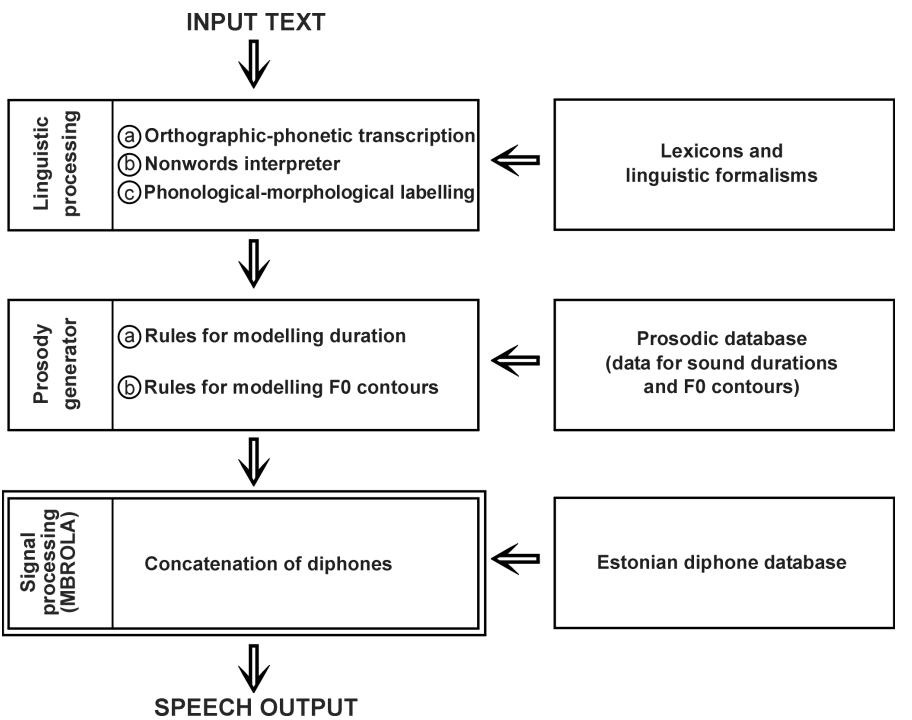


Figure 1. Block diagram of an Estonian diphone synthesizer

2.2. Synthetic voices based on unit selection

Whereas the diphone database used in the first generation compilation-based synthesis contained precisely one diphone for each sound-to-sound transition, speech synthesis based on unit selection draws from acoustic material of the whole speech corpus. The creation of a synthesizer using corpus-based unit selection is a process consisting of the following six stages:

- creation of a text corpus;
- reader selection and the storing of speech material;
- tagging of the corpus material and its segmentation into speech units;
- establishing the sound system and choice of parameters;
- morphological text analysis;
- selection of speech units for synthesis.

The text corpus needs to be sufficiently representative both from the phonetic and phonological aspects, containing all possible diphones, a great deal of numbers and years as well as frequent Estonian words and fixed combinations (Piits et al. 2007). Our minimal text corpus contains 400 sentences while the bigger ones have up to 4500. In the beginning we only used readers with trained voices (radio announcers) expected to be able to present the whole text with relatively stable prosodic parameters. However, as some of the announcers tended to display excessive expressiveness we also recorded the reading of ordinary people and in the end the synthesis based on the material of their untrained voices turned out quite satisfactory. The speech corpora are tagged automatically (Sphinx), but the output speech requires some manual finishing touches on the duration of the speech sounds adjacent to intra-sentence pauses. As Estonian is a word-central language the module of morphological analysis and disambiguation is very important for pronunciation. Unit selection and corpus-based synthesis was done (Mihkla et al. 2008) by using the Festival development system. The corpus-based synthetic voices have been derived from five donor voices:

- Riina (50 minutes of female speech)
- Tõnu (66 and 376 minutes of male speech)
- Liisi (116 and 434 minutes of female speech)
- Tõnis (71 minutes of male speech)
- Einar (89 minutes of male speech)

2.3. Formant synthesizer based on eSpeak

Estonian formant synthesis has also been realized on the eSpeak open- code development system for speech synthesis. The return to formant synthesis is mainly concerned with possible mobile applications. Besides compactness and multilinguality (over 40 languages) the eSpeak-et development version provides for speech synthesis supporting different platforms and using the markup languages SSML or HTML, and it can be used as a front-end to MBROLA diphone voices. The main shortcoming

of formant synthesis is the segmental quality of the synthesized speech, which is responsible for the machine-like sound of the output. The speech is, nevertheless, clear and intelligible enough, while the available smart interfaces enable application of an eSpeak synthesizer in almost any operation system and its use in the voice applications based on different hardware.

2.3.1. HTS speech synthesis

In statistical parametric speech synthesis based on Hidden Markov Models (HMM), speech models are trained on moderate sized corpora. Although the segmental quality of the speech wave is lower than in unit selection, the HTS method has several advantages over the latter. First, the output speech is very fluent, with an almost ideally smooth speech melody. Second, real-time process management of the synthesis provides easy access to changing the speaking rate, voice pitch and timbre. As the HTS speech synthesis descends from the clustergen synthesis supported by Festival, statistical-parametric synthesis can make use of Festvox modules and the speech corpora meant for unit selection. Several Estonian-speaking HTS voices have been synthesized.

2.4. Comparison of the synthesis modules

Table 1 compares the characteristics of Estonian speech synthesizers. The most compact of the modules is obviously eSpeak-et, which is already accessible to mobile phone applications. The synthetic voices requiring unit selection and large speech corpora are, for the time being, evidently confined to server-based applications. The best segmental quality of the output speech has been achieved by means of unit selection. The quality of speech melody (speech prosody) and smoothness is the best in HTS synthesis. As for eSpeak its options of text processing are limited. Synthesis based on unit selection denies direct access to modification of individual parameters (speech rate, voice pitch, loudness and timbre). While variability in the speech wave often causes problems for speech recognition, too little variability in speech synthesis leads to monotonous and unnatural output speech (Tatham, Morton 2005). The most variable and most natural speech is no doubt achieved by the method of unit selection.

Table 1. Comparison of the modules of Estonian speech synthesis

Module	*Compactness*	*Segmental quality*	*Smoothness of output speech*	*Level of text processing*	*Level of process control*	*Speech variability*
MBROLA-et	moderate	high	moderate	high	high	moderate
Unit selection-et	low	high	moderate	high	low	high
eSpeak-et	high	low	moderate	moderate	high	low
HTS-et	moderate	moderate	high	high	high	moderate

3. Application and challenges of speech synthesis

3.1. ELTE audio

The ELTE audio system http://elte.eki.ee/ of e-texts and audio books is designed for the visually impaired people, enabling them to read the news, papers, magazines and books and listen to audio books over the Internet. The keywords of the web system are a largest possible usership and ease of implementation, so that the system could be used not only with home PCs but also, e. g. in libraries, at public internet access points and in public transport. The audio system has been made especially easy to handle for the blind. Selection of any news report, newspaper article or book, as well as browsing of newspapers and navigation in the audio library is enabled by using just the number keys and a couple of function keys. The software has options for browsing and retrieval in the server of information literature, for choosing the rate of reading and listening as well as the voice to be synthesized. The user can choose between different modules of Estonian TTS synthesis. Both unit selection and HTS synthesis offer a menu of several male and female voices. Access to the news, papers and audio magazines is free for all. The reading of e-books and listening to audio books requires a login; the passwords are available from the Estonian Library for the Blind to all its registered readers.

3.2. Voicing of subtitles

In Estonia dubbed films are less popular than films and telecasts aired in the original language with subtitles. However, there are many people with visual impairment who can see what is going on in the screen, but cannot see well enough to read the subtitle, as well as people who cannot read for other reasons (e. g. little children or people suffering from dyslexia). Digital television can provide an additional sound canal, where sound is generated from subtitle text (see Fig. 2). Separate canals for the telecast sound and the voiced subtitles enable the user to mix the sounds of the speech synthesizer and of the TV program (receiver mix). The service is accessible in many countries, e. g. Finland. Together with the Estonian Public Broadcast we have now entered a test period during which people with visual impairment pick the best synthetic voice, while a database of pronunciation of foreign names is compiled and the whole system is tuned in.

3.3. Automatic generation of audiobooks

Next year an amendment takes effect in Estonia that obligates publishers to submit, at the same time that a hard copy is published, the text files to the National Library. The law implements the principle of equal treatment, which means that the books should be equally accessible to the sighted as well as to the blind. In collaboration with the National Library we are now developing an automatic system for audiobook generation by means of a speech synthesizer. Like in subtitle voicing considerable attention is paid to the pronunciation of foreign names and the choice of the voice to be synthesized.

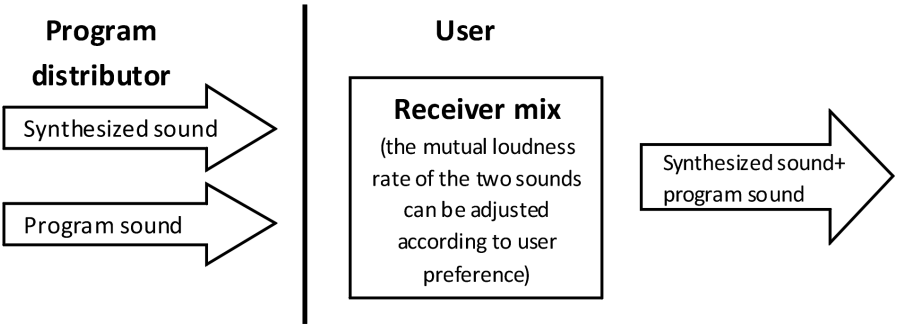


Figure 2. Receiver mix of voiced subtitles and the program sound by means of a digibox

3.4. Modelling speech prosody

One of the key concepts in speech technology is variability. Too little variability makes synthesized speech sound monotonous and unnatural. For the naturalness of the output speech we render as best possible the variation in the duration of speech sounds and pauses, the change of pitch throughout the phrase, the dynamics of speech intensity and the positions of pauses in the speech flow. In Estonian speech the temporal structure and rhythm have a special role as the phonologically significant tripartite opposition of the Estonian quantity degrees (Q1, Q2, Q3) is manifested on foot-level. In modelling quantity opposition one has to identify statistically significant features in speech temporal structure. A closer look at the phonetic parameters of Q1, Q2, Q3 in standard Estonian has revealed that specification of quantity opposition should rely on the durational relations between the vowels of stressed and unstressed syllables rather than on the durational relations between adjacent sounds (Kalvik et al. 2010). Also, the duration rate named first is a more stable parameter than pitch contours, which characterize and distinguish quantity degrees conditionally. A new approach to the higher pitch value, the position of which in the first vowel of the stressed syllable differentiates between the two long quantity degrees (Q2, Q3), suggests that comparison should be applied to the mean pitch values of different syllables (Mihkla et al. 2011). This facilitates the use of pitch (F0) as a phonetic parameter.

At word and phrase levels speech temporal structure raises various prosodic questions challenging speech synthesis. Estonian word stress is fixed on the first syllable, while the quantity opposition makes foot central for Estonian phonetics. It is still unclear whether Estonian speech rhythm is based on stress or on syllable. According to Asu, Nolan (2006) the rhythmic basis of Estonian speech might be the foot. If a sentence or phrase has a rhythmic structure of fixed feet, combined with quantity opposition, it is vital to know if and how the stress levels of the main stress syllables may change to fit the words into phrase.

One of the essential challenges facing the improvement of speech melody is to find out the relations of prosody with other language levels and extralinguistic features. If speech synthesis is based on nothing more than text phonetics, it cannot reach natural speech. Of prosodic relations, the change of the duration of the word-forming speech sounds depending on the part of speech and inflectional form of the word has been investigated. The experiments showed a few percent decrease of the output error if morphosyntactic and part-of-speech information had been added to the durational model (Mihkla 2007).

3.5. Emotional speech synthesis

Synthesis of emotional speech is one of the major challenges facing Estonian speech synthesis. At present, on the basis of the Estonian Emotional Speech Corpus http://peeter.eki.ee:5000/ (Altrov, Pajupuu 2012) the acoustic characteristics (pauses, speech rate, formants, intensity, pitch) of three basic emotions (anger, sadness and joy) and of neutral speech are being analysed. The aim is to ascertain what acoustic parameters are involved in speech emotionality and which of them are specific to this or that emotion, so that emotions could be modelled for speech synthesis. One of our principles is that speaker emotions can be identified sufficiently well from natural, i. e. non-acted speech and that non-acted speech is a prerequisite of natural speech synthesis (Iida et al. 2003). The emotions of the corpus sentences have been identified in perception tests, i. e. an emotion or neutrality has been regarded as identified if at least 51 % of the testers have agreed on it. The following four of the five acoustic parameters affecting speech emotion have been researched by now:

- pauses
- speech rate
- formants
- speech intensity

The analysis of pauses (Tamuri 2010) shows that for all three emotions concerned most of the pauses coincide with punctuation marks and there are more of breathing pauses than of non-breathing ones. According to the measurements of mean durations the longest pauses, for all three emotions, are those signalled of by a full stop and the shortest ones are the non-breathing pauses and the comma pauses. As for speech rate it was investigated if and how emotions may affect speech temporal structure. For this purpose the articulation rate of emotional utterances was measured in speech sounds per second and the results were compared with neutral speech data. Ranking the average speech rates from the most rapid to the slowest yields the following sequence: anger > joy > neutral > sadness (see Fig. 3).

In a formant analysis of emotional speech (Tamuri 2012) the first and second formants of the short stressed vowels *a*, *i* and *u* occurring in read emotional and neutral speech were measured. In pairwise analysis of the emotions (neutral included) it was found that statistically significant differences exist between the F1 means of the

vowels *a* and *i*. Intensity was measured at the midpoint of each vowel both in emotional and neutral speech, paying attention to the means and range of intensity. According to the results intensity is a significant feature in distinguishing emotions. The rank order from the highest to the lowest intensity is as follows: neutral > anger > joy > sadness.

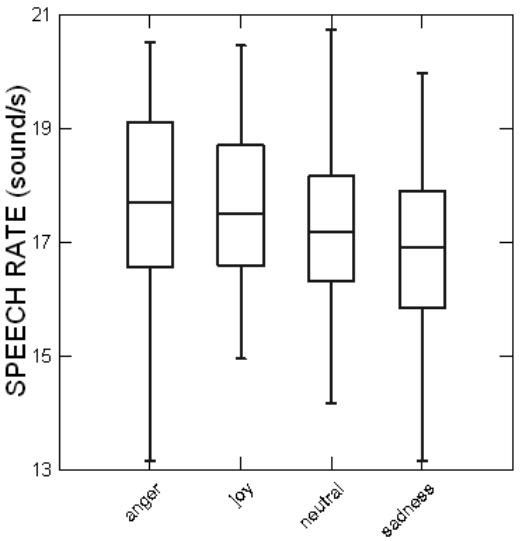


Figure 3. Correlation between speech rate and emotions

3.6. Speech synthesis interfaces

The uses of both speech synthesis and speech recognition could be enhanced if everybody interested had access not only to the modules of synthesis and recognition but also to smart interfaces (Speech Application Programming Interface). By means of such interfaces users could control the process of speech synthesis, to monitor text-to speech transformation, to follow text structure and vary the parameters (voice strength [loudness?], speech rate, voice pitch) of the synthetic voice in various voice applications. We are currently working on a development project aimed at producing prototypes of speech synthesis interfaces based on different platforms, enabling to control Estonian speech synthesis in different engines (eSpeak, Clustergen, unit selection, Mbrola, HTS), providing the widest possible coverage of the existing operation systems and hardware. For building such a SAPI (Speech Application Programming Interface) the existing speech synthesis modules (linguistic text processing, prosody generator and signal processing) are conformed to the information architecture of synthesis interfaces. Speech synthesis interfaces should guarantee that voice applications understand document structure, provide effective user navigation and

support the more widespread mark-up languages (XML, SSML). At present near-perfect speech synthesis interfaces are available only with eSpeak supporting formant and, partly, diphone synthesis of Estonian. Interfaces for other modules are still being elaborated.

4. Conclusion and prospects

In the 21st century Estonian speech synthesis has been developed using the more widespread methods and freeware development systems (MBROLA, Festival, eSpeak, HTS). The efforts have mainly been focused on the development of language-specific modules, importing methods of signal processing. The applications have hitherto been developed mainly in view of the needs of the visually impaired (audio system for reading electronic texts, voicing of subtitles, creation of audiobooks). Nowadays that speech synthesis has become part of virtual reality its potentials could be increasingly used, e. g., in language learning, computer games or information retrieval. One of our following challenges is implementation of speech synthesis in web dictionaries, so that learners of Estonian as a foreign language could hear the pronunciation of words and their inflected forms as well as of sample sentences. Hitherto a wider use of speech synthesis has been hindered by the limited expressive power of synthetic speech. Synthesis of the basic emotions belongs to the main challenges facing Estonian speech synthesis. Analysis of the emotion-bound acoustic parameters (pauses, speech rate, formants, intensity and pitch) enable one to model emotions for speech synthesis. One of the essential challenges facing the improvement of speech melody is to find out the relations of prosody with other language levels as well as with extralinguistic features. Of prosodic relations, the change of the duration of the word-forming speech sounds depending on the part of speech and inflectional form of the word has been investigated. The experiments showed a few percent decrease of the output error if morphosyntactic and part-of-speech information had been added to the durational model. More widespread use of speech synthesis could also be enhanced if everybody interested could implement any existing module of Estonian speech synthesis in their own application programs. For that purpose smart speech synthesis interfaces should be created for all of the modules, with possibly widest coverage of the existing operation systems and hardware.

5. Acknowledgements

This work was supported by the National Programme for Estonian Language Technology, grant ETF7998 and project SF0050023s09.

References

1. *Altrov R., Pajupuu H.* (2012), Estonian Emotional Speech Corpus: Content and options. G. Diani, J. Bamford, S. Cavalieri (Eds.). Variation and Change in Spoken and Written Discourse : Perspectives from Corpus Linguistics. Amsterdam : John Benjamins, [forthcoming].
2. *Asu E. L., Nolan F.* (2006), Estonian and English rhythm: a two-dimensional quantification based on syllables and feet. R. Hoffmann, H. Mixdorff (Eds.). Speech Prosody. Dresden: TUDpress, 249–252.
3. *Iida A., Campbell N., Higuchi F., Yasumura M.* (2003), A corpus-based speech synthesis system with emotion. Speech Communication 40, 161–187.
4. *Kalvik M.-L., Mihkla M.* (2010), Modelling the Temporal Structure of Estonian Speech. In: Human Language Technologies. The Baltic Perspective : Proceedings of the Fourth International Conference, Baltic HLT : Riga, Latvia, October 7–8, 2010. (Eds.) Skadina I., Vasiljevs A . Amsterdam: IOS Press, Frontiers of Artifical Intelligence and Applications, 219, 53–60.
5. *Mihkla M., Eek A., Meister E.* (1999), Diphone synthesis of Estonian. In: Dialogue'99 : Computational Linguistics and its Applications: International Workshop: Proceedings. Vol. 2. Applications. (Eds.) Narin'yani, A. S. Tarusa, 351–353.
6. *Mihkla, Meelis (2007). Modelling speech temporal structure for Estonian text-to-speech synthesis: feature selection. Trames : Journal of the Humanities and Social Sciences, 11(3), 284–298.*
7. *Mihkla M., Kalvik M.-L.* (2011), Significant features of Estonian word prosody. In: Proceedings of the 17th International Congress of Phonetic Sciences: The 17th International Congress of Phonetic Sciences (ICPhS XVII), Hong Kong, China, August 17–21, 2011. (Eds.) Wai-Sum Lee & Eric Zee. Hong Kong: City University of Hong Kong, 1378–1381.
8. *Piits L., Mihkla M., Nurk T., Kiissel I.* (2007), Designing a speech corpus for Estonian unit selection synthesis. In: Nodalida 2007 Proceedings: The 16th Nordic Conference of Computational Linguistics, 367–371.
9. *Remmel M., Tago, T.* (1984), Towards increasing the commercial success of speech synthesizers. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84.: 91–93.
10. *Taylor P.* (2009), Text-to-Speech Synthesis. Cambridge University Press.
11. *Tatham M., Morton K.* (2005), Developments in Speech Synthesis. John Wiley & Sons, Ltd.
12. *Tamuri K.* (2010), Kas pausid kannavad emotsiooni? Eesti Rakenduslingvistika Ühingu Aastaraamat, 6, 297–306.
13. *Tamuri K.* (2012), Kas formandid peegeldavad emotsioone? Eesti Rakenduslingvistika Ühingu Aastaraamat, 8, 231–243.

REPRESENTING AND REASONING FOR EXPLICIT, IMPLICIT AND IMPLICATED TEXTUAL INFORMATION

Moldovan D. (moldovan@hlt.utdallas.edu)
Human Language Technology Research Institute
University of Texas at Dallas

Explicit, implicit and implicated text meaning

It is useful to distinguish between explicit and implicit information, and between implicit and implicated knowledge conveyed by natural language.

Explicit information is what the reader gathers only from the strict meaning of words. For instance, the explicit information conveyed by *The car is cheap* (S_1) is that a particular automobile is inexpensive.

Implicit information, also called **impliciture**, is built up from the explicit content of the utterance by conceptual strengthening or "enrichment", which yields what would have been made fully explicit if lexical extensions had been included in an utterance. In our example, one may complete the sentence saying that the car is cheap compared to other cars.

Implicated information, called **implicature**, is completely separated from what is said and goes beyond what is said. They are additional propositions external to what is said, heavily dependent on the context of the situation. The speaker of S_1 may want to make a derogatory remark about the driver of the car, or may praise its owner who is rather wealthy for driving an average car and not wasting unnecessary money on a vehicle. Similarly, *Mary has a boyfriend* implicates that Mary has exactly one boyfriend. However, many other implicatures are possible depending on the context; for example: one should not ask Mary out, Mary is not a lesbian, Mary is getting a divorce, or she will be getting a divorce.

Most implicated information is conveyed during conversations. For example, given the following exchange between two users of twitter.com:

A: *Dinner's ready! prawns, grouper in some sauce, vegetables, rice and shark's fin melon soup! Still waiting for lotus root soup this week!*
B: *Eeeeeee lotus root?*
A: *so what you having for dinner?*

several facts are stated *explicitly* (*the dinner is ready*, a list of dishes: *prawns, grouper with sauce, rice, soup* made of *shark's fin* and *melon*, *lotus root soup* for later in the week, A's question about what B will have for dinner). *Implicitly*, A's first utterance discloses the existence of a *meal* which includes all the dishes it explicitly

mentions. B's *Eeeeeee* points to *disgust*, a feeling of dislike. Last but not least, the *implicatures* conveyed by each utterance complete its understanding tying together the explicit and implicit information it communicated: *A has prepared the ready dinner which includes the list of mentioned dishes*; *A is excited of having prepared this gourmet dinner, B dislikes lotus root and cannot believe that A would choose to eat it*; *A has a poor opinion of B's gastronomic knowledge*.

We note that text understanding ranges from extracting explicit information to more complex implicitures or entailments, all the way to implicatures that take into account contextual information. Thus, implicitures or implications are implicit in what is said, whereas implicatures are implied by what is said. Implicatures depend largely on context of the utterance, whereas implications depend less of context. Sentences imply certain things, but speakers implicate. Speaker utters a sentence and in fact implicate something else which can be true or false.

Hierarchical knowledge representation

In order to represent the information conveyed by natural language texts (explicit, implicit or implicated), we proposed a hierarchical knowledge representation, consisting as several layers (Figure 1). Each layer in the hierarchy builds upon the previous layer by bringing new information. The base is the *lexico-syntactic* layer, which reveals the text's syntactic parse trees. The *semantic* layer adds semantic relations between the concepts identified by the previous layer. Next, the identified *events* bring a new level of abstraction to the process of representing knowledge since events tend to dominate the meaning of text. Last, *relations between events* are established with the goal of determining the text coherence at the highest level.

In order to derive the explicit information conveyed by natural language, which is to be represented using the formalism described above and stored for later use by more advanced applications, a library of robust NLP tools are needed to extract lexical, syntactic and semantic information from natural language input. Our NLP pipeline includes a text tokenization module, a part-of-speech tagging system, a sentence boundary detection tool, a collocation identification module, a named entity recognizer, a chunk parser, a word sense disambiguation system[4], a full-syntactic parser [1], a nominal and pronominal coreference resolution module, a semantic parser [2],[9], and an event and event relation identification tool [3],[5].

Logical representation of knowledge

In order to facilitate the consumption of derived knowledge by a natural language reasoner, the hierarchical information must be converted into a *logical form*. For this purpose, a logical predicate is created for each textual concept that participates in a semantic relation. The name of each predicate derived from a concept is a concatenation of the lexeme's base form, its part of speech, and corresponding WordNet sense number. For questions, a special predicate is introduced to capture the

target type, the answer type of the question. Furthermore, for each predicate associated with a named entity or event, a new predicate is added to the logical form of the text that describes this property. Last, but not least, logical predicates denoting the semantic relations and event relations identified during the NLU process are included to complete the logical representation of the derived knowledge.

Given these simple logical transformation of text rules, an accurate multi-layered first order logical form of the knowledge conveyed by *Gilda Flores's kidnapping occurred on January 13, 1990. A week before, he had fired the kidnappers* can be derived (Figure 1).

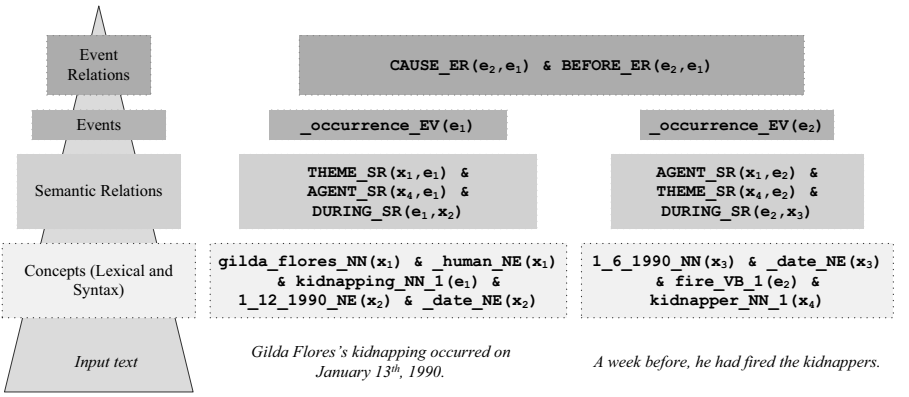


Figure 1. Hierarchical knowledge representation

More specifically, all concepts mentioned in the text constitute the bottom layer of the representation. Named entity classes (e. g., *Gilda Flores* = *human*) and normalized values of temporal expressions (e. g., *a week before January 13, 1990* = 01/06/1990) are captured within this layer of lexical information. The next level of knowledge is dedicated to semantic relations identified between the text's concepts (e. g., theme(*Gilda Flores*, *kidnapping*)). Co-referring concepts are also stored in this layer of representation (*he* = *Gilda Flores*). Events and their properties are identified and represented as the third layer of information (e. g., *kidnap*, theme(*Gilda Flores*), agent(*kidnappers*), during(*01/13/1990*)), with the final level of representation consisting of event-event relations (cause(*fire, kidnapping*), before(*fire, kidnapping*)). Finally, there is a causation relations, namely, firing (e_2) is the cause of kidnapping (e_1) and has preceded the kidnapping. The reason for kidnapping is the firing of employees and the two sentences can be summarized as a *revenge* action.

A model for representing knowledge

In order to represent the meaning of natural language, we define a model that captures not only the *explicit* information transmitted to the hearer, but also the *implicit* and *implicated* information conveyed by a speaker.

By characterizing a speaker S by his utterance ($S_U = \{u_1,..., u_m\}$, where u_i is a speaker utterance (what speaker says)) and his intentions ($S_I = \{i_1,..., i_k\}$, where i_i is a speaker intention conveyed by S_U) and the hearer H by his understanding of S_U (H_{TM} denotes the hearer understanding of the *explicit* text meaning; H_{TE} represents the hearer understanding of the *implicit* text meaning, i. e. entailments, and H_I symbolizes *implicatures* H made from S_U), the 7-tuple

$$\{S_U, S_I, H_{TM}, H_{TE}, H_I, C, K\}$$

captures the *complete* meaning of an utterance. Our model's *C* component represents the context and *K* denotes the common sense knowledge needed by H to derive S's implicatures.

Noting that each of these components are hierarchical representations of explicit, implicit as well as implicated knowledge conveyed by a natural language text, the model establishes a standard of representation for natural language meaning.

An instantiation of this model for A's first utterance, part of the exchange shown above, *Dinner's ready! prawns, grouper in some sauce, vegetables, rice and shark's fin melon soup! Still waiting for lotus root soup this week!* is

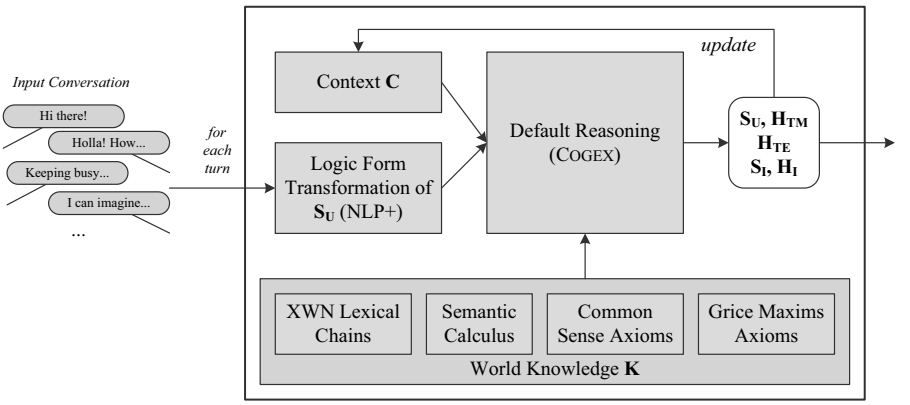
S_U/H_{TM}: DINNER_NN_1(x_1) & READY_JJ_1(x_2) & VALUE_SR(x_2,x_1) & EXCLAMATORY_TYPE & PRAWN_NN_1(x_3) & GROUPER_NN_1(x_4) & SAUCE_NN_1(x_5) & PART_WHOLE_SR(x_4,x_5) & VEGETABLE_NN_1(x_6) & RICE_NN_1(x_7) & SHARK_NN_1(x_8) & FIN_NN_1(x_9) & PART_WHOLE_SR(x_9,x_8) & MELON_NN_1(x_{10}) & SOUP_NN_1(x_{11}) & PART_WHOLE_SR(x_{10},x_{11}) & PART_WHOLE_SR(x_9,x_{11}) & STILL_RB_1(x_{12}) & WAIT_VB_1(e_1) & _OCCURRENCE_EV(e_1) & MANNER_SR(x_{12},e_1) & LOTUS_NN_1(x_{13}) & ROOT_NN_1(x_{14}) & PART_WHOLE_SR(x_{14},x_{13}) & SOUP_NN_1(x_{15}) & PART_WHOLE_SR(x_{14},x_{15}) & THEME_SR(x_{15},e_1) & WEEK_NN_1(x_{16}) & _DATE_NE(x_{16}) & DURING_SR(e_1,x_{16})
H_{TE}: DISH_NN_1(x_{17}) & PART_WHOLE_SR(x_5,x_{17}) & MEAL_NN_1(x_{18}) & PART_WHOLE_SR(x_{17},x_{18}) & PART_WHOLE_SR(x_5,x_{18}) & DISH_NN_1(x_{19}) & ISA_SR(x_{11},x_{19}) & MEAL_NN_1(x_{20}) & PART_WHOLE_SR(x_{19},x_{20}) & PART_WHOLE_SR(x_{11},x_{20})
S_I/H_I: SHOW_VB_1(e_2) & _A_USER(x_{21}) & AGENT_SR(x_{21},e_2) & STRONG_JJ_1(x_{22}) & FEELING_NN_1(x_{23}) & VALUE_SR(x_{22},x_{23}) & THEME_SR(x_{23},e_2) & =(x_1,x_{18},x_{20}) & COOK_VB_1(e_3) & THEME_SR(x_{18},e_3) & AGENT_SR(x_{21},e_3)
C: ∅

Reasoning

The NLU methodologies mentioned above are the first step taken by automated systems to determine the *complete* meaning of a natural language text. They derive the *explicit* information conveyed by the text: the meaning of the words used by the speaker (given by the named entity recognizer or the word sense disambiguation modules) and their semantic connections (given by the coreference resolution module

and the semantic parser). However, in order to go beyond what is being said, to derive the *implicit* and *implicated* information conveyed by the text, and, thus, to expand the understanding of explicit information, reasoning mechanisms that exploit the derived knowledge using various types of natural language axioms within an abductive reasoning framework must be implemented.

The natural language reasoner presented here processes the natural language of a conversation one turn at a time and derives the logical implications and implicatures communicated by each speaker utterance, updating the conversation's context after each individual analysis. Below, we show the architecture of the system, highlighting the roles and interactions of our model's components ($\{S_U, S_I, H_{TM}, H_{TE}, H_I, C, K\}$).



Cogex, a natural language reasoning engine

Cogex is a heavily modified version of the Otter theorem prover[1], which uses the Set of Support (SoS) strategy to prove by contradiction: a hypothesis is proved by showing that it is impossible for it to be false in the face of the provided evidence and background knowledge (BackgroundKnowledge, Evidence & ¬Hypothesis → ^).

Cogex has been successfully employed within a question answering engine to rerank the final list of candidate answer passages based on the degree of entailment between each passage and the given question [16]. For the Recognizing Textual Entailment task, Cogex computes the extent to which a text snippet entails a hypothesis as a normalized score between 0 and 1 and compares this value to a threshold learned during training to determine whether the given hypothesis is entailed by the given text [6],[7],[8].

Within these settings, the initial usable list contains various natural language axioms, which can be used to infer new information (BackgroundKnowledge), while the logical clauses corresponding to the candidate passage/text snippet (Evidence)

[1] http://www.cs.unm.edu/~mccune/otter/

as well as the negated question/hypothesis (¬Hypothesis) are added to the SoS list. Given Otter's best-first clause selection mechanism, the heavily weighted question/hypothesis clauses are the last clauses to be processed by the system. Thus, when resolutions using these clauses are attempted, it is guaranteed that all other inferences ([BackgroundKnowledge & Evidence]+) have been made and are stored in the usable list.

For the task of deriving *implicit* and *implicated* information conveyed by natural language, Cogex exploits the first order logical representations of the meaning model components. It makes use of the S_U, C, and K components to derive the values of S_I , H_{TM}, H_{TE}, and H_I , which, in turn, will be used to update the context of future conversational turns (Figure 2). Given that our current assumptions include that the communication channel is noise-free, $H_{TM} = S_U$. Futhermore, $S_I = H_I$ since it is difficult to determine whether the speaker intended all the implicatures S_U conveyed at the time of S_U's analysis. These values are revised if future conversational turns cancel some of the implicatures derived during S_U's analysis.

When the analysis of a conversation begins, the context C is empty, the set of axioms described above form the knowledge component (K), and the semantic representation of the speaker utterance make up the model's S_U component. Furthermore, the first order clauses of the C and K model components form the usable list and the logical clauses of S_U serve as the initial SoS. As the reasoning process unfolds, all S_U predicates and the inferences they produce (entailments as well as implicatures) are moved to the usable list, where they become part of the context C of future conversational turns.

The reasoning process terminates when no more inferences can be made from S_U (i. e., SoS is empty). In this situation, the clauses inferred from a given utterance using non-default axioms and/or context clauses that were explicitly stated or entailed by previously analyzed utterances (i. e., previous S_U/H_{TM} and H_{TE} components) are marked as logical entailments, or *implicitures* (H_{TE}). All clauses inferred from default axioms as well as abductive rules and/or context clauses that were previously labeled as implicatures (i. e., previous S_I/H_I components) are marked as the *implicatures* conveyed by the current S_U (S_I,H_I). This is the most expected outcome of an analysis of a conversational turn.

However, if a refutation is found during the reasoning process, the clauses that caused the inconsistency may indicate that (1) the speaker made a false statement (the contradiction stems from S_U clauses alone), which carries certain Quality-flouting implicatures or (2) a previously identified implicature must be canceled (information from current S_U contradicts previous S_I/H_I), in which case, the implicature clauses are removed and the reasoning process is restarted.

Natural language axioms

The axioms used during the reasoning process capture the world knowledge people use to derive the information *implicitly* conveyed by natural language texts. Various types of axioms are needed to encode this vast pool of information.

eXtended WordNet lexical chain axioms

The lexical chain axioms link WordNet concept by exploiting the semantic relationships present in the eXtended WordNet (XWN)[2]. In addition to WordNet's semantic relations between synsets, this valuable semantic resource stores a highly semantic logical representation of WordNet's plain text glosses, a rich source of world knowledge made available to automated systems by our NLP tools. Using the structured representation of WordNet's glosses provided by XWN, new semantic links are identified between WordNet's synsets. Examples include:

$\text{PRAISE_VB_1}(e_1)$ & $\text{AGENT_SR}(x_1, e_2)$ & $\text{THEME_SR}(x_2, e_1)$ → $\text{EXPRESS_VB_1}(e_2)$ & $\text{ISA_SR}(e_1, e_2)$ & $\text{APPROVAL_NN_1}(x_3)$ & $\text{THEME_SR}(x_3, e_2)$ & $\text{AGENT_SR}(x_1, e_2)$ & $\text{THEME_SR}(x_2, x_3)$ — WordNet gloss for *praise*: "express approval of"

$\text{PHONE_CALL_NN_1}(x_1)$ → $\text{TELEPHONE_NN_1}(x_2)$ & $\text{ISA_SR}(x_1, x_2)$

Semantic calculus axioms

We introduced the notion of Semantic Calculus to describe the various combinations of two semantic relations. The Semantic Calculus axioms describe the semantic relation R_0 that holds between two concepts c_1 and c_2 linked by two semantic relationships R_1 and R_2 (not necessarily distinct) that share a third concept c_3 as a common argument. More formally, $R_1(c_1,c_3)$ & $R_2(c_3,c_2) \rightarrow R_0(c_1,c_2)$ [12],[13].

We note that not any two semantic relations can be combined. Furthermore, the accuracy of each Semantic Calculus axiom was measured as the correctness with which it predicted the correct semantic relation that links two concepts also connected by the semantic relations that are combined (for each linguistic context where $R_1(c_1,c_3)$ and $R_2(c_3,c_2)$ hold, for any c_1, c_2, and c_3, consider if $R_0(c_1,c_2)$ also holds).

These axioms greatly increase the semantic connectivity between concepts. This is particularly important when no immediate semantic link can be found between two concepts of interest. Examples include:

$\text{ISA_SR}(x_1, x_2)$ & $\text{ISA_SR}(x_2, x_3)$ → $\text{ISA_SR}(x_1, x_3)$

$\text{ISA_SR}(x_1, x_2)$ & $\text{LOCATION_SR}(x_2, x_3)$:→ $\text{LOCATION_SR}(x_1, x_3)$

$\text{QUANTITY_SR}(x_1, x_2)$ & $\text{INSTRUMENT_SR}(x_3, x_1)$:→ $\text{QUANTITY_SR}(x_1, x_3)$

Common sense axioms

This type of axioms encodes the common sense knowledge required by an automated system to derive unstated implications. These axioms describe not only various

[2] http://xwn.hlt.utdallas.edu/index.html

properties of concepts, but also how the concepts interact in the world and how people speak about them. Most of the semantic rules belonging to this class are default axioms. They also have an abductive nature, implicating the best-guess outcomes. Fully automatic methods for deriving these semantic axioms are yet to be proposed. Examples of world knowledge axioms include:

TOPIC_SR(x_1,x_2) & PURPOSE_SR(e_1,x_2) & COMMUNICATE_VB_1(e_1) → TOPIC_SR(x_1,e_1) — this axiom states that *the topic of something used for communication becomes the topic of the communication*

EXPRESS_VB_1(e_1) & APPROVAL_NN_1(x_1) & THEME_SR(x_1,e_1) & PERSON_NN_1(x_2) & THEME_SR(x_2,x_1) & AGENT_SR(x_2,e_2) :→ THEME_SR(e_2,x_1) — this axiom states that *if expressing approval of person that is the agent of action, then express approval of action*

Furthermore, valuable world knowledge can be identified by exploring the semantic relations identified in textual documents. Unlike the semantic relations stored within machine readable dictionaries, such as WordNet, these relationships link concepts mentioned within a document's content that belong to all part-of-speech classes. These semantic relations are contextual, locally true, and may not seem useful for a reasoning engine that operates outside of their context. However, most of the concepts they connect are defined in WordNet's hierarchies and, thus, they can be abstracted to more general concepts. Using this process of generalization, we can axiomatically describe the selectional restrictions of certain concepts by noting regularities in the types of arguments that they pair with.

For instance, the axiom shown below can be derived based on the high frequency of agent semantic relations that link the verb *communicate* and various hyponyms of *person* — we include here all *_human* named entities, most of which are not included in WordNet:

COMMUNICATE_VB_1(e_1) :→ _HUMAN_NE(x_1) & AGENT_SR(x_1,e_1) — this axiom states that *people are usually the agents of communicating*

We note that these semantic rules are default axioms. Each is associated with a weight that is proportional with the frequency of the defining semantic relation within a large corpus.

Grice Maxims axioms

The implicit assumptions and default inferences that capture our intuitions about a normal interpretation of a communication have been analyzed and formalized by language philosophers as principles of rational human communication behavior. For instance, Grice's Cooperative Principle explains the link between utterances and what is understood from them. Grice's Maxims [14],[15], which are derived from this principle, can be leveraged by an automated system aiming to derive the *implicatures* conveyed by natural language utterances.

The set of macro-axioms encoding the essence of Grice's Maxims make use of the components of the model we defined above. Before being used by an automated system for the analysis of a particular utterance, each Gricean maxim axiom must be instantiated with the values of the various components of the existing conversational model. Examples include:

$_\text{RELEVANCE}_\text{GM} \rightarrow (\text{PREDICATE}_i(x_j) \in \text{LF}(S_U)\ \&\ \text{PREDICATE}_i(x_k) \in (C) :\rightarrow x_j = x_k)$ — this axiom states that *there must be at least one common predicate between the (enhanced) logical forms of* S_U (speaker utterance) *and* C (context) — given that the speaker's utterance must be relevant to the established common ground. In the case of floutings of the Relevance Maxim, this axiom can be used to *assume the unification of two identically named predicates from* S_U *and* C

$_\text{QUALITY}_\text{GM}\ \&\ S_U(x_1)\ \&\ _\text{QUESTION}(x_1) \rightarrow S(x_2)\ \&\ \text{BELIEVE}_\text{VB}_1(e_1)\ \&\ \text{EXPERIENCER}_\text{SR}(x_2, e_1)\ \&\ H(x_3)\ \&\ \text{KNOW}_\text{VB}_1(e_2)\ \&\ \text{EXPERIENCER}_\text{SR}(x_3, e_2)\ \&\ \text{THEME}_\text{SR}(e_2, e_1)\ \&\ \text{THEME}_\text{SR}(x_1, e_2)$ — this axiom states that for all speaker utterances that are questions, S (the speaker) *believes* H (the hearer) *knows the answer to his question*

$_\text{MANNER}_\text{GM} \rightarrow ((_\text{CAPITALIZED}_\text{CHARS}(x_1)\ \&\ S_U(x_1)) :\rightarrow (S(x_2)\ \&\ \text{EXCITED}_\text{JJ}_1(x_3)\ \&\ \text{VALUE}_\text{SR}(x_3, x_2)))$ — this axiom states that *capitalized texts within the speaker's utterance indicate the excitement of the speaker*

$_\text{QUANTITY}_\text{GM}\ \&\ S_U : (P(x_i) \rightarrow P(x_i)) \rightarrow \text{PREDICATE}_0(x_0) \in \text{LF}(C)\ \&\ (\text{PROPERTY}_\text{SR}(x_0, x_i)\ |\ \text{VALUE}_\text{SR}(x_0, x_i))\ \&\ _\text{ALWAYS}_\text{TMP}(x_0)\ \&\ -(\text{PREDICATE}_x(x_x) \rightarrow -(\text{PROPERTY}_\text{SR}(x_0, x_i)\ |\ \text{VALUE}_\text{SR}(x_0, x_i)))$ — this axiom states that *if the speaker's utterance is of the form "X is X", then one of X's (essential) properties that is relevant to the existing common ground always happens in X and nothing can change that*

References

1. *Elliot Glaysher and Dan Moldovan*. 2006. Speeding Up Full Syntactic Parsing by Leveraging Partial Parsing Decisions. In Proceedings of COLING/ACL 2006, Sydney, Australia, July 2006
2. *Adriana Badulescu and Munirathnam Srikanth*. 2007. LCC-SRN: LCC's SRN System for SemEval 2007 Task 4. In Proceedings of SemEval-2007, Prague, Czech Republic, June 23–24, 2007
3. *Congmin Min, Munirathnam Srikanth, and Abraham Fowler*. 2007. LCC-TE: A Hybrid Approach to Temporal Relation Identification in News Text. In Proceedings of SemEval-2007, Prague, Czech Republic, June 23–24, 2007
4. *Adrian Novischi, Muirathnam Srikanth, and Andrew Bennett*. 2007. LCC-WSD: System Description for English Coarse Grained All Words Task at SemEval 2007. In Proceedings of SemEval-2007, Prague, Czech Republic, June 23–24, 2007
5. *Marta Tatu and Munirathnam Srikanth*. 2008. Experiments with Reasoning for Temporal Relations between Events. In Proceedings of COLING 2008, Manchester, UK, 18–22 August, 2008

6. *Marta Tatu, Brandon Iles, John Slavick, Adrian Novischi and Dan Moldovan*. 2006. COGEX at the Second Recognizing Textual Entailment Challenge. In Proceedings of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment. Venice, Italy, May 2006
7. *Marta Tatu and Dan Moldovan*. 2006. A Logic-based Semantic Approach to Recognizing Textual Entailment. In Proceedings of COLING/ACL 2006, Sydney, Australia, July 2006
8. *Marta Tatu and Dan Moldovan*. 2007. COGEX at RTE 3. In Proceedings of ACL-2007 Workshop on Textual Entailment and Paraphrasing. Prague, Czech Republic, June 2007
9. *Dan Moldovan and Eduardo Blanco*. 2012. Polaris: Lymba's Semantic Parser. In Proceedings of LREC 2012, Istanbul, Turkey, May 2012
10. *Dan Moldovan and Adriana Badulescu*. 2005. A Semantic Scattering Model for the Automatic Interpretation of Genitives. In Proceedings of HLT-EMNLP 2005, Vancouver, British Columbia, Canada, October, 2005
11. *Dan Moldovan and Vasile Rus*. 2001. Logic Form Transformation of WordNet and its Applicability to Question Answering. In Proceedings of ACL-2001, Toulouse, France, July 2001.
12. *Eduardo Blanco and Dan Moldovan*. 2011. Unsupervised Learning of Semantic Relation Composition. In Proceedings of ACL-HLT 2011, Portland, OR, USA.
13. *Eduardo Blanco and Dan Moldovan*. 2011. A Model for Composing Semantic Relations. In Proceedings of IWCS 2011, Oxford, UK.
14. *Grice H. P.* 1969. Utterer's Meaning and Intentions. In Philosophical Review 78: 147–177 [Reprinted in [15]]
15. *Grice H. P.* 1989. Studies in the Way of Words. Cambridge, MA: Harvard University Press
16. *Dan Moldovan, Marta Tatu, and Christine Clark*. 2010. Role of Semantics in Question Answering. In Semantic Computing, pages 373–419. John Wiley & Sons, Inc.

WHAT DO PEOPLE ASK THE COMMUNITY QUESTION ANSWERING SERVICES AND HOW DO THEY DO IT IN RUSSIAN?

Mukhin M. (mfly@sky.ru)
Ural Federal University, Yekaterinburg, Russia

Braslavski P. (pbras@yandex.ru)
Kontur Labs/Ural Federal University, Yekaterinburg, Russia

In our study we surveyed different approaches to the study of questions in traditional linguistics, question answering (QA), and, recently, in community question answering (CQA). We adapted a functional-semantic classification scheme for CQA data and manually labeled 2,000 questions in Russian originating from Otvety@Mail.Ru CQA service. About half of them are purely conversational and do not aim at obtaining actual information. In the subset of meaningful questions the major classes are requests for recommendations, or how-questions, and fact-seeking questions. The data demonstrate a variety of interrogative sentences as well as a host of formally non-interrogative expressions with the meaning of questions and requests. The observations can be of interest both for linguistics and for practical applications.

Keywords: community question answering, interrogative sentence, question types, Otvety@Mail.Ru

Introduction

Community question answering (CQA) is a popular on-line social activity. CQA sites allow users to pose questions to other community members, to answer questions, rate questions and answers, get scores, etc. Yahoo! Answers[1], Answers.Com[2] and Otvety@Mail.Ru[3] are examples of popular general-purpose CQA services. Stackoverflow[4] is an example of a domain-specific CQA service which specializes in software programming. Quora[5] represents a newer type of such service where questions and answers can be updated, followed, interlinked, etc., thus generating potentially higher-quality content and making it more reusable.

[1] http://answers.yahoo.com/

[2] http://www.answers.com/

[3] http://otvet.mail.ru/

[4] http://stackoverflow.com/

[5] http://www.quora.com/

CQA became a good complement to Web search engines: they satisfy users' complex information needs, find answers to opinionated questions and questions that imply practical experience and accounting for context. To date CQA services have collected a vast amount of data: for example, Yahoo! Answers claimed reaching one billion questions & answers in October 2009.[6] On the one hand, CQA data (not only textual, but also user activity and interaction data) help improve existing services, rethink question answering (QA) and build value-added services on top of the collected data. On the other hand, the data are a valuable linguistic resource where researchers get access to a large amount of living language material from millions of informants that is partially structured (question — list of answers) and categorized by topics. Although many CQA services look alike and some services are operated globally, the usage patterns can vary in different countries, influenced by local traditions and culture as [24] showed. Thus, analysis of language- and country-specific data is important.

In our study we develop a framework for classification of Russian-language questions originating from a popular CQA service and manually classify 2,000 questions. This task is of interest both for linguistics and for practical applications. To the best of our knowledge, this is the first study of language material of the sort.

The rest of the paper is organized as follows. The following section surveys the literature on linguistic approaches to questions and the work on classification of questions in context of QA and CQA. Section 3 briefly describes Otvety@Mail.Ru service and the data used in our study. In Section 4 we introduce a framework for classification of CQA questions; Section 5 summarizes the results of manual classification of 2,000 questions. The sixth section contains conclusions and outlines directions for future research.

Related Work

In this section we survey the three groups of related work: 1) classification approaches to interrogative sentences and questions in linguistics, 2) question classification approaches in question answering (QA), and 3) question typologies introduced within CQA research.

Question as a semantic category and interrogative sentence are asymmetric phenomena; in the semantic description of the two it is crucial to set "a clear distinction between interrogative sentence as a syntactic notion and question as a semantic category required by the information structure" [17: 233]. Most linguistic studies focus on interrogative sentences rather than questions as such. However, formal and semantic aspects are often inseparable in traditional classifications. Thus, classical studies consider interrogative sentence among the other types of sentences of different purposes of communication (such as declarative, imperative, and, in some classifications, optative sentences), which "serves to express the question posed to the other party. With the help of the question the speaker seeks to obtain new information about something..." [11: 302]. Similar definitions can be found in [7, 19], and others.

[6] http://yanswersblog.com/index.php/archives/2009/10/05/did-you-know/

Question as a type of statement with a particular communicative task — that of inducement to obtain information [5: 707] — can be structured both in the form of an interrogative and a non-interrogative sentence. By its nature, it can be a request, a demand, etc. On the other hand, interrogative sentence in its primary function may or may not express the speaker's desire to obtain new information, i. e., to be "properly interrogative" or "improperly interrogative" [5: 708], and to have "standard" or "non-standard" interrogative semantics [17: 233–234][7]. Based on this, Bulygina and Shmelev proposed the two issues for linguistic consideration: "1) how questions are expressed (besides interrogative sentences), and 2) what function interrogative sentences fulfill (other than their primary function of expressing questions)" [5: 111]. The second problem is reflected in a large number of studies. As for the former issue, we think that social services on the web may help to make a decision on it. They are characterized by large amounts of data, different ways of information search, informal register of communication, along with the necessary restrictions on the dialogue, which are missing in regular online forums. Among these relatively new linguistic data we are primarily interested in the functional-semantic question types (not interrogative sentences as such), as well as in characteristics of potential answers embedded in the question.

Automatic classification of questions is an important problem in the area of question answering (QA). QA is a subfield of information retrieval (IR), where user information need is formulated as a natural language question (rather than a list of keywords), resulting in an exact answer or a concise document fragment containing the answer — in contrast to a ranked list of documents in a classical IR scenario. This direction of research has largely shaped and demonstrated progress thanks to the TREC QA track (see http://trec.nist.gov/data/qa.html, [22]). The main type of questions used for QA evaluation within TREC was open-domain factual questions, or factoids, e. g. *What was the monetary value of the Nobel Peace Prize in 1989?* (Later, questions seeking for definitions and relational information were added.) The overall performance of a QA system is heavily influenced by the ability of the system to predict the type of the expected answer based on the question. TREC participants used a wide range of question typologies and classification approaches that were tightly connected both with the TREC data and the named entity recognition (NER) output. The collection of about 5,500 labeled questions known as the UIUC dataset became a de-facto standard in the field ([12], http://cogcomp.cs.illinois.edu/Data/QA/QC/). The proposed hierarchy contains six coarse classes and 50 fine classes, see Fig. 1.

Coarse	**Fine**
ABBREVIATION	abbreviation, expansion
ENTITY	definition, description, manner, reason

[7] From this point of view classifications of interrogative sentences (categorized as direct and indirect speech acts) are presented, besides those already mentioned works, in [3, 5, 6, 13, 20, 25], and others.

Coarse	Fine
DESCRIPTION	animal, body, color, creation, currency, disease/medical, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
HUMAN	description, group, individual, title
LOCATION	city, country, mountain, other, state
NUMERIC VALUE	code, count, date, distance, money, order, other, percent, period, speed, temperature, size, weight

Figure 1. UIUC dataset question typology [13]

Several attempts have been made to enrich the UIUC typology or to tailor it to a particular task. For example, a recent study [16] analyzes questions in Korean, including those from a search engine query log and proposes a classification scheme based on the three facets:

- Answer format (AF): *single (factoid)*, *multiple (list)*, *descriptive (definition)*, and *yes/no*.
- Answer theme (AT, similar to UIUC types) is the class of the object sought by the question, such as *person*, *location*, or *date*. A total of 147 themes are organized in a hierarchy and are derived from a NER task.
- Question qualifier (QQ) reflects a question's semantics or pragmatics; the possible values are *specification*, *superlative*, *ordering*, *definition*, etc.

CQA has recently attracted attention of researchers from different fields — information retrieval, linguistics, as well as sociology and related disciplines. [1] gives a good insight into the nature of CQA services and analyzes various aspects and characteristics of Yahoo! Answers: the differences and similarities among categories, the activity of users, their interactions, etc. CQA sites contain a vast amount of user generated content (UGC) that significantly varies in quality. [2] addresses the problem of automatic identification of high-quality questions and answers in a dataset obtained from Yahoo! Answers. Using a wide range of features — content features, usage statistics, and user relationships — the authors were able to separate high-quality items from the rest with high accuracy.

Several studies have been done on classification of questions asked on CQA services. [8] introduced the question dichotomy *conversational* vs. *informational*, where the former questions are asked purely to start discussion and the latter are aimed at satisfying an actual information need. About 500 questions from different CQA services were annotated manually according to the scheme. Then, the authors implemented a classifier based on category, question text and asker's social network characteristics. [14] investigated a similar facet of Q&A threads, namely *social* vs. *non-social* intent of the users: all questions intended for purely social engagement are considered social, while those that seek information or advice are considered non-social but instigating a knowledge sharing engagement. 4,000 questions from two different CQA services were labeled.

Harper et al. later proposed a rhetorical question typology consisting of the three type pairs [9]: Advice and Identification, (Dis)Approval and Quality, and Prescriptive and Factual. Each pair is considered to be a subspecies of Aristotelian rhetorical genres: deliberative, epideictic, and forensic, respectively. 300 Yahoo! Answers questions were labeled manually according to this typology.

[10] adopted a psycholinguistic typology for labeling about 800 questions from Yahoo! Answers focused on the following topics: data mining, natural language processing (NLP), and eLearning. The annotation scheme consisted of the nine types: Concept Completion, Definition, Procedural, Comparison, Disjunctive, Verification, Quantification, Causal, and General Information Need.

[15] investigated how people ask and answer questions in online social networks (primarily on Facebook and Twitter). The authors classified 249 questions provided by survey participants into the following categories: Recommendation, Opinion, Factual knowledge, Rhetorical, Invitation, Favor, Social connection, and Offer.

Data

Otvety@Mail.Ru (*otvety* means *answers* in Russian) is a service of Mail.Ru, one of the leading Russian web portals. Otvety@Mail.Ru is a Russian counterpart of Yahoo! Answers, with similar rules and incentives (see Fig. 2). It was launched in August 2006, and after five years has reached 50M+ users, 60M+ questions, and 335M+ answers. The service claims to have 58K new users, 52K questions, and 235K answers daily.[8]

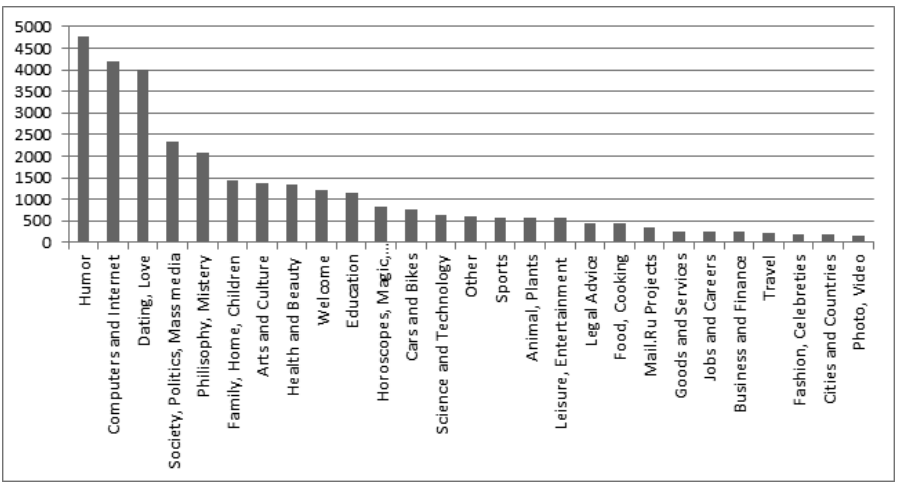


Figure 2. Otvety@Mail.Ru user interface

[8] http://otvet.mail.ru/news/#hbd2011

An Otvety@Mail.Ru question consists of a title (often it is the question itself, up to 120 characters), a detailed question description (optional) that may contain links, images, and video along with the text; tags (optional), category and subcategory (mandatory, are chosen manually by the asker from the drop-down lists).

For our initial experiments we downloaded every 1,000th question and its answers for the period from September 2009 to November 2010. It resulted in 31,223 non-empty pages. Fig. 3 shows the distribution of pages across the top-level categories. The average question length (concatenation of question title and optional question body) in our dataset is 22.5 words, while the average answer length is 19.7 words; a question receives 5.3 answers in average.

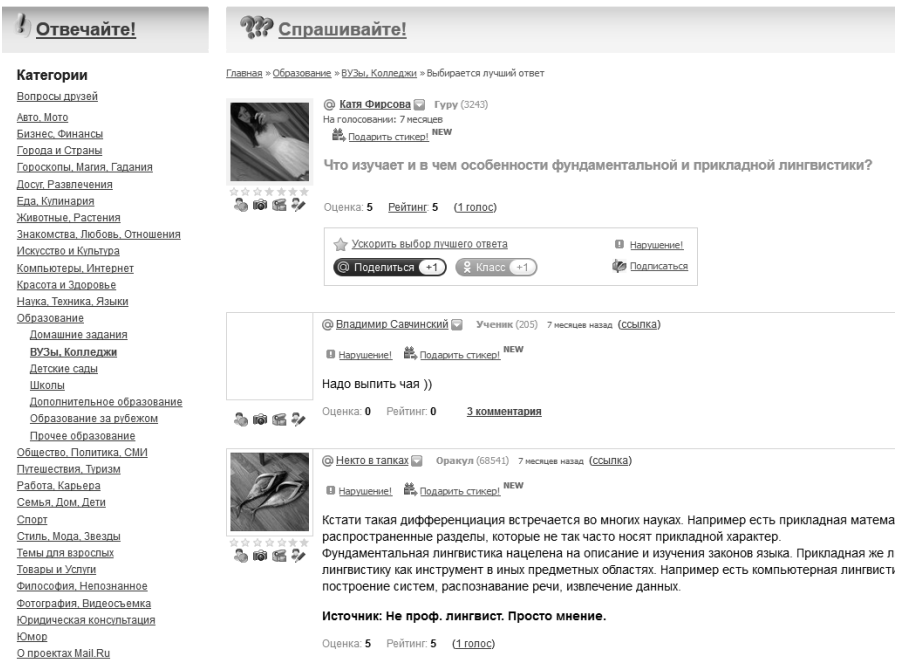


Figure 3. Distribution of Q&A pages by category

An initial examination of the data allowed us to make some observations that imply a broad understanding of the question category in CQA.

1. A large number of "questions" with non-standard semantics do not relate to information search *per se*, but are rather invitations to conversation or an opportunity for the asker to express herself (to make a joke, to shock others, etc.), i. e. they carry only a secondary function. Examples: *Подойдет ли монтажная пена для макияжа!?))) [Is foam sealant suitable for makeup!?)))]; Как вы думаете будет ли такое время что все люди будут жить в единстве в мире и любви? [Do you think there will be a time when all people will live in unity, peace and love?]*

2. Many formally non-interrogative structures can be paraphrased and represented as traditional questions, for instance: *антивирус не обновляется из-за ошибки компилятора*; *Выбираю авто [the antivirus does not update due to compiler's error; Choosing auto mode]; Opel Astra, многие её хвалят, но хотелось бы узнать ваше мнение.[Opel Astra, many praise it, would like to know your opinion]*

3. Several questions that can imply different types of responses are combined together, e. g.: *Является ли философия наукой? И если да, то почему в ней так слабо развит математический аппарат?[Is philosophy a science? If yes, why is its mathematical apparatus so underdeveloped?]*

4. Different uses of question title and body fields:

- Question topic in the title, detailed question in the body (well-formed structure): *Проблемы с Windows 7 || Как запускаю игру, мой ноут просто выключается. В чем проблема? [Problems with Windows 7|| When starting the game my laptop turns off. What is the problem?]*; *Капитанская дочка || Почему Маша, любя Гринёва, отказывается выйти за него замуж?! [The Captain's Daughter[9]||Why does Masha, being in love with Grinev, refuse to marry him?]*;
- The title contains the question, while the body is either empty, or contains a clarification or a request for help or answer: *Жёсткий диск какой фирмы посоветуете приобрести? || Баракуду не предлагать [What manufacturer's hard drive would you recommend?||Do not suggest Barracuda]; А кто солил арбузы? || Киньте рецептик, коль не жалко <…> [Who had made pickled watermelons? || Drop in a receipt, please…]*;
- The title contains an appeal for help, an address, or the beginning of an answer: *ответьте на вопросы плиззз))) || Для чего Князь Андрей отправляется на войну [Answer the question pleazzze))) || What for does Prince Andrew go to war[10]]; Дорогие хозяйки своего очага, подскажите, пожалуйста, || как и чем вывести жирное пятно (от крема) с дивана (шинил)? [Dear housewives, tell me, please, || how to remove a greasy spot (cream) from a sofa (chenille)?]*;
- The title and the body contain different, even if related questions: *Какова вероятность забеременеть сразу, если был незаметный выкидыш? || И что назначают врачи после выкидыша? [What are the odds of becoming pregnant shortly after a minor miscarriage? || And what do doctors prescribe after a miscarriage?]*;
- The question body merely repeats the title: *Что подарить девушке на 17 лет?[What is a good present for a girl's 17th birthday?]*.

5. The wide topical range of questions spans from requests to help in solving a math problem or a crossword puzzle to requests for legal advice or a desired link, etc.

---

[9] A novel by Alexander Pushkin, studied in middle school.

[10] The question relates to the novel "War and Peace" by Leo Tolstoy, also studied in the school.

Question typology

When analyzing Otvety@Mail.Ru data we relied both on traditional approaches to question classification, and on recent work on questions in online social networks and on search queries in the form of questions (see Section 2). The proposed question classification is functional-semantic by its nature, generalizing the substantial characteristics of questions rather than those of interrogative sentences.

The nature of the data implies iterative refinement of the classification (union or, in contrast, subdivision of classes, similarly to [14]) and suggests the following types of questions.

According to the main function we can distinguish the following classes:

1) Actual questions seeking for information, in broad terms: it can be traditional questions, requests for help or advice, as well as invitations to join a community, to make use of something (in this case we are talking about real, concrete facts, events, or matters);

2) *Rhetorical* questions and remarks that do not ask for information, 'chat' — the same as *conversational* in [8] and *rhetorical* in [15]: it can be an invitation to a conversation (even on a serious topic), a joke or an emotional expression:

Вы один из тех, каких много или считаете себя особенным?))) [Are you one of those who consider themselves extraordinary?];

Есть ведро солёных огурцов. Сколько надо вёдер водки?[There is a bucket of pickles. How many buckets of vodka are needed?]

Explicit questions can be further characterized by their particular functions, defined through the expected answer type or the action of the potential interlocutor. For our study we adopted the classification scheme from [15] as the top-level categories and elaborated a finer-grained layer for the two major classes.

1. *Factual knowledge* — the search for factual information. This category is further divided into the following subclasses:

- Object (*Как называется…?*[11], *Кто/Что это? [What/Who is…?]*);
- Object property (*Чем отличается…?*; *Как выглядит…?[How does X look…?]*; *Как действует…?[How does X work…?]*);
- Possibility (*Можно ли…?[Is it possible…?]*, *Могу ли я…? [Can I…?]*, i. e. a question that asks about possibility/impossibility of doing something);
- Reason (*Почему это так…? [Why…?]*, *О чем говорит…?[What does it mean…?]* — about the objective reason of a property or event);
- Aim (*Зачем нужен…?*, *Для чего…? [What for…?]*);
- Time (*Во сколько…?*, *Когда произойдет…? [When…? At what time will X happen?]*).

[11] Here we consider both formal interrogative constructions and statements that can be reformulated as a question. For example, a question of type "Object" can be simply a link to an online image.

2. *Recommendation* — a question or request of the type «*please tell me, how to…*». A further specification is the following:

- Method (literally *Как сделать…? [How to make…?]*);
- Information search (predominantly on-line — *Где найти…?[Where can I find…?], Как узнать?[How can I know…?]* etc.);
- Location, directions (*Куда поехать?, Где находится…? [Where is X located…? How can I get to…?]*— in the geographic sense).

3. *Opinion — Как вы относитесь к…? [What do you think about…?], Что вы предпочитаете?[What do you prefer…?].*

4. *Favor* — asking for help: *Пришлите ссылку… [Please send a link…], Решите задачу… [Please solve a problem…],* etc.

5. *Offer* (*Кому нужен… [Who needs…?], Отдам/Продам…[…to sell/…to give away]*).

6. *Social connections* — the search for people, business companions, friendship, love and sexual relations.

We also complemented the hierarchy above with the expected answer type following [16] (see Section 2).

Results

2,000 randomly sampled questions from the dataset were labeled manually by one of the authors according to the proposed scheme (Section 4). What follows is a summary of the main results.

993 cases (49.7%) were assigned to the *rhetorical* class ('chat'); 1007 (50.3%) were actual questions seeking for meaningful information. A further division of the question class is presented in Table 1. As one can see, the majority of the questions are seeking for procedural knowledge (*Recommendation*) or facts (*Factual Knowledge*); are asking for a favor (*Favor*) or inquire *Opinions* of others. Many questions of the *Favor* type are connected with acquisition of procedural knowledge (e. g. asking for help with a math assignment), thus we can conclude that the questions that can be reformulated as *How to make something?* prevail at Otvety@Mail.Ru. Questions that imply social interaction (*Offer* and *Social Connection*) are presented marginally.

Table 1. Subclasses of actual question class (total 1007)

Question type	Count	%
1. Factual knowledge	343	34.1
2. Recommendation	391	38.8
3. Opinion	123	12.2
4. Favor	135	13.4
5. Offer	4	0.4
6. Social connection	11	1.1

Table 2. Subdivision of two major question classes

	Count	%
1. Factual knowledge	**343**	**100**
1.1. Object	119	34.7
1.2. Object property	135	39.4
1.3. Possibility	41	12.0
1.4. Reason	35	10.2
1.5. Aim	7	2.0
1.6. Time	6	1.7
2. Recommendation	**391**	**100**
2.1. Method	251	64.2
2.2. Information search	120	30.7
2.3. Location, directions	20	5.1

The data in Table 2 are fairly predictable, as factual questions deal mostly with objects and their properties. Note the presence of the subgroup *Possibility* with such questions as whether it is possible to file a court petition or, for example, to buy a new SIM card.

Table 3. Breakdown of the sample by expected answer type

Answer type	Count	%
Yes/no	66	6.6
Single	241	23.9
Multiple	166	16.5
Descriptive	534	53.0

Table 3 shows that the majority of questions imply a detailed answer (description), i.e. a simple statement of a fact would not be sufficient. In contrast, there are very few yes/no questions (*Is the temperature of 46 degrees OK for an Intel Core 2 processor? Can I have a badger at home?*). As Table 4 reveals, the majority of the *Recommendation* questions imply a detailed response. A portion of the fact-seeking questions expects yes/no; however, short and descriptive answers prevail.

Table 4. Distribution of questions by expected answer type in two major classes

Answer type	Factual knowledge (total 343)	%	Recommendation (total 391)	%
Yes/no	44	12.8	1	0.3
Single	118	34.4	61	15.6
Multiple	40	11.7	88	22.5
Descriptive	141	41.1	241	61.6

Conclusions

In our study we refined and combined the question classification schemes reported in the literature. 2,000 questions in Russian originating from Otvety@Mail.Ru CQA service were tagged manually. The dataset is one of the biggest manually tagged collections of CQA questions reported in the literature (only the dataset reported in [14] excels in size) and the first Russian-language collection of this kind, to the best of our knowledge. About half of the questions is aimed rather at self-expression, joking and chatting than seeking for information and knowledge sharing. The portion of the "entertaining" questions seems to be higher in Otvety@Mail.Ru than in other similar CQA services according to [8, 14] (however, all the studies including ours use different classification schemes, so any comparisons should be done carefully). Obviously, this proportion of meaningful and conversational questions is determined by many features and characteristics of a CQA service: the broad audience, its social and demographic characteristics, absence of topical focus, the system of incentives, and the moderation policy.

Questions on CQA services demonstrate a different behavior compared to question-like status messages on Facebook and Twitter [15] — for instance, there are far fewer rhetorical and fact-seeking questions. The reason could be that in online social networks users can satisfy their needs in socializing and conversations without resorting to question-like messages; users seek chiefly for recommendations and opinions from their immediate social network.

It is interesting to note that the prevalence of *recommendations*, or *how* questions in our data mirrors the trend in search engine query logs: the *how* question-like queries surpass other pragmatics [18, 23]. The situation has changed a lot during the last decade: in the late 90s most search engine queries in question form sought for factual information [21].

The processed data demonstrate a variety of interrogative sentences as well as of formally non-interrogative expressions with meaning of questions and requests. The obtained data can be of interest for linguists, sociologists, and communication researchers. The data can also be used for improving existing CQA services and can contribute to question answering research. The manually tagged sub-corpus of questions is available for research purposes at http://kansas.ru/cqa/data/.

In our future research we are going to address the problem of automatic classification of questions, refine our classification scheme, as well as compare CQA questions with questions in search engine query logs.

We would like to thank Mail.Ru for granting us access to the data and Tanya Kondakova for initial data preparation.

References

1. *Adamic L. A., Zhang J., Bakshy E., and Ackerman M. S.* Knowledge Sharing and Yahoo! Answers: Everyone Knows Something. Proceedings of the 17th International Conference on World Wide Web (WWW '08). ACM, New York, NY, USA, 2008, pp. 665–674.
2. *Agichtein E., Castillo C., Donato D., Gionis A., Mishne G.* Finding High Quality Content in Social Media. Proceedings of the International Conference on Web Search and Web Data Mining (WSDM '08). ACM, New York, NY, USA, 2008, pp. 183–194.
3. *Bally Ch.* Linguistique générale et linguistique française [Obshchaja lingvistika i voprosy francuzskogo jazyka]. Moscow, Inostrannaya Literatura, 1955, p. 416.
4. *Beloshapkova V. A.* Types of Sentences for Aim [Tipy predlozhenij po celeustanovke]. Sovremennyj russkij jazyk: Uchebnik dlja vuzov [The Modern Russian Language: Textbook for Universities]. Edited by V. A. Beloshapkova, 3rd ed. Moscow, Azbukovnik, 1999, pp. 705–708.
5. *Bulygina T. V., Shmelev A. D.* Interrogative sentences and Questions [Voprositel'nye predlozhenija i voprosy]. Chelovecheskij faktor v jazyke: Kommunikacija, modal'nost', dejksis [The Human Factor in Language: Communication, Modality, Deixis]. Moscow, Nauka, 1992, pp. 111–116.
6. *Conrad R.* Interrogative Sentences as Indirect Speech Acts [Voprositel'nye predlozhenija kak kosvennye rechevye akty]. Novoe v zarubezhnoj lingvistike. Vyp. 16: Lingvisticheskaja pragmatika [Novelty in Foreign Linguistics: Linguistic Pragmatics]. Vol. 16. Moscow, Progress, 1985, pp. 349–383.
7. *Gvozdev A. N.* The Modern Russian Literary Language: Syntax: Part 2 [Sovremennyj russkij literaturnyj jazyk: Sintaksis: Ch. 2]. 5th ed. Moscow, Librokom, 2009. 352 p.
8. *Harper F. M., Moy D., and Konstan J. A.* Facts or Friends? Distinguishing Informational and Conversational Questions in Social Q&A Sites. Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI '09). ACM, New York, NY, USA, 2009, pp. 759–768.
9. *Harper F. M., Weinberg J., Logie J., and Konstan J.* (2010), "Question types in social Q&A sites". First Monday [Online], Volume 15 Number 7 (4 July 2010), available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2913/2571
10. *Ignatova K., Toprak C., Bernhard D., and Gurevych I.* Annotating question types in social Q&A sites. Tagungsband des GSCL Symposiums "Sprachtechnologie und eHumanities" Abteilung für Informatik und Angewandte Kognitionswissenschaft Fakultät für Ingenieurwissenschaften Universität Duisburg-Essen, 2009, pp. 44–49.
11. *Lekant P. A.* Interrogative Sentences [Voprositel'nye predlozhenija]. Sovremennyj russkij literaturnyj jazyk: Uchebnik dlja vuzov [The Modern Russian Literary Language: Textbook for Universities]. Edited by P. A. Lekant, 4th ed. Moscow, Vysshaja Shkola, 1999, pp. 302–305.
12. *Li X., Roth D.* Learning Question Classifiers. Proceedings of the 19th International Conference on Computational Linguistics — Volume 1 (COLING '02),

Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, pp. 1–7.
13. *Lindstrem E. N.* Classification of Russian Interrogative in Form Statements on the Basis of a Pragmatic-Based Universal Model [Klassifikacija russkih voprositel'nyh po forme vyskazyvanij na baze pragmaticheski obosnovannoj universal'noj modeli]. PhD diss. Petrozavodsk, 2003.
14. *Mendes-Rodrigues E. and Milic-Frayling N.* Socializing or Knowledge Sharing? Characterizing Social Intent in Community Question Answering. Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09). ACM, New York, NY, USA, 2009, pp. 1127–1136.
15. *Morris M. R., Teevan J., and Panovich K.* What Do People Ask Their Social Networks, And Why?: A Survey Study Of Status Message Q&A Behavior. Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI '10). ACM, New York, NY, USA, 2010, pp. 1739–1748.
16. *Oh H.-J., Sung K.-Y., Jang M.-G., Myaeng S. H.* (2011) Compositional Question Answering: A Divide and Conquer Approach. Information Processing & Management, Vol. 47, Issue 6, pp. 808–824.
17. *Paducheva E. V.* The Statement and Its Correlation with Reality [Vyskazyvanie i ego sootnesennost' s dejstvitel'nost'ju]. Moscow, Nauka, 1985. 271 p.
18. *Pang B., Kumar K.* Search in the Lost Sense of "Query": Question Formulation in Web Search Queries and its Temporal Changes. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers — Volume 2 (HLT '11), Vol. 2. Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 135–140.
19. *Rozental' D. E., Telenkova M. A.* Dictionary of Linguistic Terms [Slovar'-spravochnik lingvisticheskih terminov]. Moscow, Prosveshchenie, 1976.
20. *Russian* Grammar [Russkaja Grammatika]. Moscow, Nauka, 1980. Vol. 2.
21. *Spink A., Ozmultu H. C.* (2002) Characteristics of Question Format Web Queries: An Exploratory Study. Information Processing & Management, Vol. 38, Issue 4, pp. 453–471.
22. *Voorhees E. M., Dang H. T.* Overview of the TREC 2005 Question Answering Track. Proceedings of TREC-14, Gaithersburg, MD, USA, 2005.
23. *Yandex* (2010), "Questions to Yandex" (in Russian), available at: http://company.yandex.ru/researches/figures/ya_questions_2010.xml
24. *Yang J., Morris M. R., Teevan J., Adamic L. A., and Ackerman M. S.* Culture Matters: A Survey Study of Social Q&A Behavior. International AAAI Conference on Weblogs and Social Media, North America, 2011. Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2755.
25. *Zhinkin N. I.* Question and Interrogative Sentence [Vopros i voprositel'noe predlozhenie]. Jazyk — Rech' — Tvorchestvo. Issledovanija po semiotike, psiholingvistike, pojetike [Language — Speech — Creativity. Research on Semiotics, Psycholinguistics, Poetics]. Moscow, Labirint, 1998, pp. 87–103.

ПРИМЕНЕНИЕ N-ГРАММ И ДРУГИХ СТАТИСТИК УРОВНЯ СИМВОЛОВ И СЛОВ ДЛЯ СЕМАНТИЧЕСКОЙ КЛАССИФИКАЦИИ НЕЗНАКОМЫХ СОБСТВЕННЫХ ИМЕН

Нехай И. В. (nekhayiv@gmail.com)

Кафедра распознавания изображений и обработки текста, ФИВТ МФТИ, Москва, Россия

Автоматическая семантическая классификация неизвестных собственных имен в текстах является значимой проблемой при разработке средств автоматического анализа и обработки текста. В данной работе мы исследуем возможности классификации часто используемых в русском языке собственных имен при помощи машинного обучения и побуквенных n-грамм, а также других признаков уровня букв и слов, таких как капитализация слов, общее число слов, наличие в имени аббревиатур, чисел и т.п. Мы используем только внутренние признаки, однако, поскольку использование внешних признаков ожидается в будущих работах, мы тестируем также сценарии, которые могут оказаться полезными при использовании внешних признаков. В работе показано, что при помощи простых признаков и средств классификации можно достичь достоверности 80–99 % для биномиальной классификации, 89–99 % при отделении одной категории от остальных и 85 % достоверности для задачи мультиномиальной классификации. При этом используется обучающая выборка, содержащая порядка 2000 примеров в каждой категории. Аналогичные результаты для английского языка были продемонстрированы в работе (Patel & Smarr, 2001), где применялся аналогичный подход.[1]
Работа выполнена в рамках исследований, проводящихся под эгидой проекта Министерства Образования и Науки № 13.G25.31.0088.

Ключевые слова: n-граммы, семантическая классификация, имена собственные, именованные сущности

Введение

При разработке систем автоматического анализа и перевода текстов, использующих словари, необходимо решать проблемы, связанные с обработкой собственных имён. Не представляется возможности исчерпывающе перечислить в словарях собственные имена, такие как антропонимы, топонимы, названия компаний, фильмов, лекарств. Это вызвано следующими двумя причинами:

[1] Доклад публикуется в сокращенном варианте. Полная версия доступна на сайте конференции «Диалог».

1. на текущий момент существует огромное число подобных имён;
2. непрерывно порождаются новые имена.

Таким образом, неизвестные системе разбора текста слова необходимо обрабатывать автоматически. Существует ряд известных процедур, выполняемых при такой обработке:

1. выделение собственных имён в тексте, в частности:
   1.1. детектирование присутствия имени и
   1.2. определение границ имён;
2. приведение к нормальной форме;
3. семантическая классификация имён.

Пример реализации процедуры 1 можно найти в (Baluja, 2000), процедуры 2 — в (Mikheev, 1997).

В данной работе мы рассматриваем вариант реализации процедуры семантической классификации имен, т. е. отнесения их к какой-либо из заранее заданных нами категорий. При этом предполагается, что предшествующие процедуры уже выполнены, и входными данными для системы классификации являются выделенные и приведённые к нормальной форме имена.

Основной целью данной работы является получение программы-классификатора, использующего внутренние признаки имён собственных для классификации их по следующим категориям: антропонимы, названия городов, фильмов, компаний, лекарств. В статье (Patel & Smarr, 2001) описана сходная реализация классификатора англоязычных имён, использующего 4 категории, поэтому дополнительной целью является сравнение характеристик полученных классификаторов. Отметим особо, что система классификации исследовалась на примерах, широко распространённых в современном русском языке.

1. Основные принципы

Классификатор разработан по принципу «обучения с учителем»: программе предоставляются «обучающий» набор имён, на основе которого вычисляются статистические параметры модели, которые в дальнейшем используются для классификации «неизвестных» имён.

В процессе разработки были решены следующие вопросы:

- составление набора «обучающих» данных;
- определение набора признаков, на основании которых осуществляются обучение и классификация;
- выбор алгоритма классификации;
- определение и вычисление метрик качества классификации;

Эти вопросы рассматриваются в последующих разделах.

2. Внешние и внутренние признаки

Как отмечается в (McDonald, 1993), для классификации собственных имен можно использовать два вида признаков.

1. Внешние признаки — морфосинтаксическая информация о словоизменении имени и его роли в тексте или предложении.
2. Внутренние признаки — лексическая информация, содержащаяся в самом имени. Примерами их для английского языка, согласно (Baluja, 2000), являются:
   а. капитализация слов;
   б. написание слов заглавными буквами;
   в. наличие чисел;
   г. количество цифробуквенных последовательностей;
   д. наличие слов, состоящих из одной буквы.

В данной работе внешние признаки не принимаются во внимание, так как рассматриваются уже выделенные и лемматизированные имена, поэтому используются только внутренние признаки.

Основным внутренним признаком для классификации, как и в (Patel & Smarr, 2001), является статистика символьных n-грамм. На этапе обучения по принципу максимального правдоподобия эмпирически оцениваются вероятности $P(l_i|c_i,l_{i-k}\ldots l_{i-2}\,l_{i-1})$ того факта, что символ l_i в именах категории c встречается после последовательности символов $l_{i-k}\ldots l_{i-2}\,l_{i-1}$ длины k. На этапе классификации вероятность того, что имя w принадлежит категории c оценивается как

$$P(w|c) = \prod_{i=1\ldots len(w)} P(l_i|c_i,l_{i-k}\ldots l_{i-2}\,l_{i-1}) \qquad (1)$$

При этом значение k, определяющее глубину учитываемого контекста, является параметром и выбирается, исходя из максимизации достоверности классификации. Для корректного учёта статистики по начальным и конечным символам имени в начало имени приписывается k условных символов “$”, а в конец — один символ “#”.

Значения вероятности сглаживаются при помощи правила Лидстоуна:

$$\hat{P}(l_i|c, l_{i-k}\ \ldots\ l_{i-2}l_{i-1}) = \frac{N_{l_i}+\lambda}{N_{l_{i-k}\ldots l_{i-2}l_{i-1}}+N_l\cdot\lambda}, \qquad (2)$$

где N_{l_i} — число появлений символа l_i после последовательности $l_{i-k}\ldots l_{i-2}\,l_{i-1}$, $N_{l_{i-k}\ldots l_{i-2}\,l_{i-1}}$- число появлений этой последовательности, N_l — число уникальных символов во всех примерах, λ — параметр сглаживания.

Значение параметра λ выбрано равным 0.15, так как значительное число экспериментов показали, что это значение обеспечивает максимальную достоверность классификатора. Экспериментальная зависимость достоверности классификатора от значения λ приведена на рис. 1. Это значение также может выбираться при помощи кросс-валидации.

Остальные признаки были выбраны путём ручного анализа основных классов ошибок классификатора. В результате были получены следующие признаки (в скобках **полужирным шрифтом** отмечены используемые в графиках сокращения):

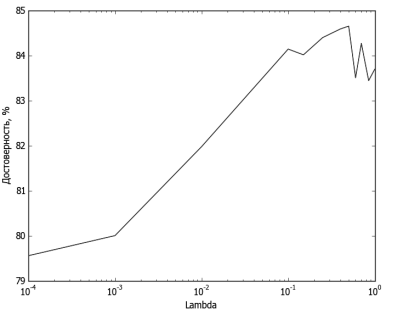


Рис. 1. Зависимость достоверности классификатора от значения параметра сглаживания λ, задача мультиномиальной классификации

- *бинарные, значения которых есть «истина» или «ложь»*:
  - капитализация всех слов имени (**ВсеКап**);
  - наличие чисел в имени (**СодЧисл**);
  - наличие аббревиатур (последовательностей длины 3 и более заглавных букв) (**СодАббр**);
  - наличие слов, содержащих заглавные буквы после строчных (пример: *НоваТЭК*) (**СлитКап**);
  - имя состоит из слов, разделённых дефисами (пример: *Ростов-на-Дону*) (**Дефисоид**);
  - имя полностью написано заглавными буквами (**ВсеЗагл**);
- *числовые, значения которых есть натуральные числа:*
  - число последовательностей цифр или букв (**ЧислоПослед**);
  - общее число символов (**ЧислоСимв**);
- *числовые, значения которых есть произвольные вещественные числа:*
  - средняя длина последовательности цифр или букв (**СредДлПослед**);
  - средняя длина слов (последовательностей букв) (**СредДлСлов**).

3. Алгоритм классификатора

В основе классификатора лежит машина опорных векторов (SVM), использующая в качестве ядра радиальную базисную функцию. Входными значениями признаков для неё являются:

- значения $\log_2 P(w|c)$ вероятностей принадлежности имени w классу c, вычисляемые по формуле (1);
- значения остальных признаков.

Заметим, что наш классификатор не использует при обучении априорные вероятности классифицируемых категорий. Это ослабляет ограничения на использование классификатора, т. е. полученные оценки справедливы для случая, когда об априорных вероятностях ничего не известно и они считаются одинаковыми, т. е. $P(c) = \frac{1}{N_c}$ для всех категорий c, где N_c — общее число всех категорий. В таком случае значения $P(w|c)$ можно использовать в качестве значений $P(w|c)$, поскольку $P(c)$ не зависит от c, и

$$argmax_c P(w|c)P(c) = argmax_c P(c|w) \quad (3)$$

Однако, если априорные вероятности категорий можно некоторым образом оценить, то они могут быть использованы путём следующего изменения формулы (1):

$$argmax_c P(c|w) = argmax_c P(c) \prod_{i=1\ldots len(w)} P(l_i|c, l_{i-k}\ldots l_{i-2}\, l_{i-1}) \quad (4)$$

Процесс обучения построен следующим образом. 80% примеров из обучающей выборки используются для оценки вероятностей n-грамм. После этого по оставшимся 20% примерам обучается машина опорных векторов с вышеперечисленными входными значениями. При этом все входные значения нормируются к отрезку [–1,1] и исключаются линейно зависимые значения. Далее статистика n-грамм перестраивается по всем примерам обучающей выборки.

Тем самым, машина опорных векторов в некоторой степени компенсирует ошибки, возникающие при классификации на основе n-грамм, т. к. примеры, на которых она обучается, не использовались для сбора статистики и отражают характер входных данных, к которым будет применяться классификатор.

4. Способы применения

В данной работе не рассматривается использование внешних признаков для семантической классификации имён. Однако мы планируем использовать полученный классификатор в рамках более сложной системы обработки текста, в которой внешние признаки также будут доступны для классификации. Поэтому мы будем осуществлять оценку результатов, исходя из различных способов применения классификатора при наличии внешних признаков.

Используемый метод опорных векторов осуществляет биномиальную классификацию, но может быть преобразован к мультиномиальной (многозначной) путём попарной классификации между всеми имеющимися категориями и выбора категории с самой высокой оценкой.

В более сложной системе анализа текста при помощи внешних признаков можно ограничить множество рассматриваемых категорий для каждого анализируемого имени, тем самым повышая достоверность классификации на основе внутренних признаков.

Таким образом, мы будем оценивать классификатор при решении следующих задач:

- биномиальная классификация имени в рамках произвольной пары категорий — частный случай задачи с уточнением в рамках заранее выбранной пары категорий;
- отделение одной категории от остальных — при решении задач выделения собственных имён определённой категории из текстов;

- мультиномиальная классификация — определение категории имени из полного списка возможных категорий.

Отметим, что возможны также задачи классификации в рамках тройки, четвёрки и т. п. категорий, но изложение экспериментальных результатов по ним было бы весьма объёмным, а никакой новой информации эти результаты не открывают, т. к. все такие задачи сводимы к перечисленным.

5. Методы оценки результатов

Для оценки результатов работы применяется оценка достоверности классификатора (в англоязычной литературе — *accuracy*), определяемая как:

$$accuracy = \frac{\text{число верно классифицированных примеров}}{\text{общее число примеров}} \quad (5).$$

Все полученные результаты приводятся в сравнении с работой (Patel & Smarr, 2001) в тех случаях, когда доступны результаты для сравнения.

Отметим, что в работе (Patel & Smarr, 2001) контрольная выборка содержала различные доли примеров из каждой категории, в связи с чем достоверность константного или произвольного классификаторов в ней может составлять до 85 % для отдельных категорий и режимов работы. В настоящей работе во всех контрольных выборках доля примеров каждой категории составляет $^{1}/_{\text{число категорий}}$, поэтому достоверность константного классификатора составляет:

- 50 % — в задачах биномиальной классификации;
- 80 % — в задачах выделения одной категории (классификатор, определяющий все примеры как не принадлежащие заданной категории);
- 20 % — в задаче мультиномиальной классификации.

6. Источники и характеристики экспериментальных данных

В качестве экспериментальных данных использовались примеры, полученные из различных интернет-источников.

В категории «Имена людей» использовались списки, содержащие имена и фамилии сотрудников компании кафедры распознавания изображений и обработки текста ФИВТ МФТИ, а также имена, фамилии и отчества депутатов Государственной Думы РФ V и VI созывов. К спискам случайным образом применялись 10 распространённых правил сокращения и перестановки имён, фамилий и отчеств, что позволяет получить практически неограниченное число правдоподобных имён. Общее число примеров данной категории, использованное в экспериментах — 2500.

В категории «Названия компаний» использовались списки рейтинга Эксперт-400 (http://www.expert.ru/) и других рейтингов российских компаний,

доступных в Интернете. Отметим, что в полученных списках отсутствуют обозначения организационно-правовой формы компании и другие признаки (кавычки и т. п.), которые существенно упрощают классификацию имён как названий компаний. Объём категории — 2509 названий.

В категории «Фильмы» использовались материалы сайта http://www.kinokadr.ru/allfilms/year/2010/ — названия фильмов, вышедших в прокат в России в период 2000–2010 гг., очищенные от пометок «DVD», «сериал» и т. п. Объём — 2713 названий.

В категории «Лекарства» использованы материалы сайта http://leim.ru/Practica/Drugs/druglist.htm, включающие как торговые марки лекарств, так и названия действующих компонентов. Объём — 1984 примера.

В категории «Названия городов» использованы материалы сайта http://autotravel.ru/towns.php, включающие 1621 название.

Примеры имён указанных категорий приведены в таблице 1.

Таблица 1. Примеры экспериментальных наборов данных

Категория	Примеры
Имена людей	*Бочкарев Андрей Валерьевич; Г. Г. Семёнов; Муцоев, Александр Генрихович; Антонов Е. А.*
Названия компаний	*Аптечная сеть 36,6; Волжская ТГК (ТГК-7); ТАИФ-НК; Авиакомпания «Домодедовские авиалинии»*
Фильмы	*Ледниковый период 4: Континентальный дрейф; Эволюция; 11.11.11; Сумерки 4: Рассвет. Часть 1*
Лекарства	*Адапален; АКДС/вакцина полиомиелитная инактивированная; Этидронат натрия; Бензокаин; Берлинская лазурь*
Названия городов	*Бережная Дуброва; Интерпосёлок; Юрьев-Польский; Ыб; Умет; Урочище Калбак-Таш*

7. Экспериментальные результаты

7.1. Биномиальная классификация

Таблица 2. Характеристики достоверности для режима биномиальной классификации

Категории	Достоверность	Эквивалентные категории (Patel & Smarr, 2001)	Достоверность (Patel & Smarr, 2001)
Люди — лекарства	99,87%		
Люди — фильмы	99,70%		

Категории	Достоверность	Эквивалентные категории (Patel & Smarr, 2001)	Достоверность (Patel & Smarr, 2001)
Люди — компании	99,35%		
Люди — города	99,36%		
Лекарства — города	96,34%	Drug — Place	94,406%
Фильмы — лекарства	96,46%	Drug — Movie	94,875%
Лекарства — компании	94,32%	Drug — NYSE	98,881%
Фильмы — города	87,26%	Movie — Place	88,955%
Фильмы — компании	80,39%	NYSE — Movie	98,887%
Города — компании	82,80%	NYSE — Place	98,721%
Средние значения	**93,585%**		**95,787%**
Средние значения без категории «Люди»	**89,595%**		**95,787%**

7.2. Задача выделения одной категории

Таблица 3. Характеристики достоверности для режима выделения одной категории

Категория	Достоверность	Эквивалентная категория (Patel & Smarr, 2001)	Достоверность (Patel & Smarr, 2001)
Люди	99,36%		
Фильмы	91,15%	Movie — Others	91,987%
Лекарства	96,82%	Drug — Others	95,722%
Города	87,39%	Place — Others	90,325%
Компании	88,98%	NYSE — Others	99,952%
Средние значения	**92,74%**		**94,321%**
Средние значения без категории «Люди»	**91,085%**		**94,321%**

7.3. Задача мультиномиальной классификации

Общая достоверность классификатора — 84–85%, при этом контрольная выборка содержит одинаковое число примеров каждой из 5 категорий.

По данным (Patel & Smarr, 2001) — 88,971%, при этом контрольная выборка содержит различное число примеров из 4 категорий.

7.4. Детализация экспериментальных результатов

В экспериментах использовалась статистика символьных n-грамм длины 3, поскольку было проверено, что именно эта длина обеспечивает максимальную достоверность классификатора (см. рис. 2).

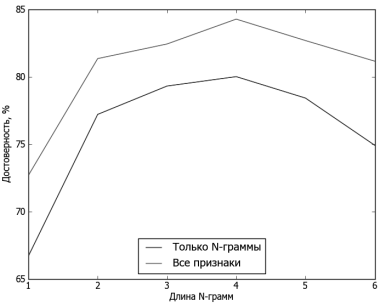


Рис. 2. Зависимость достоверности классификатора от длины используемых n-грамм, мультиномиальная классификация

Были проверены классифицирующие способности каждого признака в отдельности (рис. 3). Поскольку, согласно рис. 2, статистика n-грамм без дополнительных признаков обеспечивает достоверность порядка 80%, проверены также достоверности классификатора на основе сочетания n-грамм и одного дополнительного признака (рис. 4). На (рис. 4) также приведены для сравнения результаты классификаторов, использующих все признаки кроме N-грамм, и только N-граммы без дополнительных признаков.

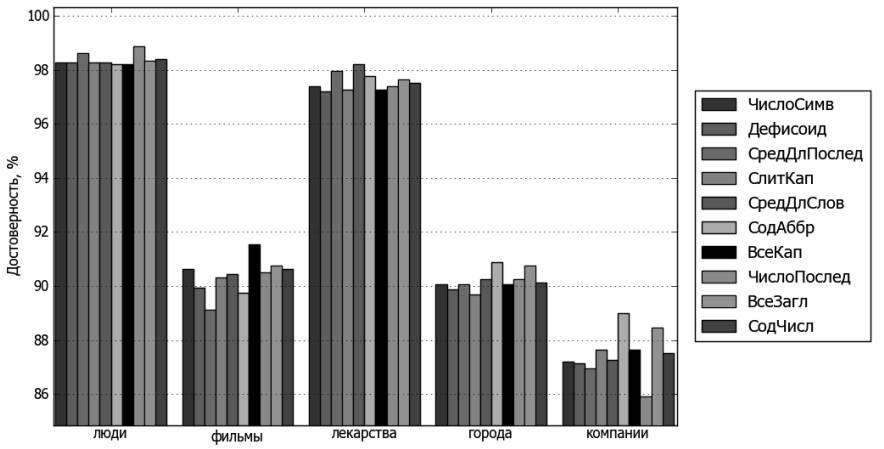


Рис. 3. Классифицирующие способности отдельных признаков, показатель достоверности, задача выделения одной категории

Подчеркнём, что достоверность 80% на рис. 3 соответствует достоверности константного классификатора (не относящего к выбираемой категории ни одного примера), что свидетельствует о полной бесполезности признака для той или иной категории. Достоверности меньшего, чем 80%, уровня говорят о том, что признак позволяет отделить от требуемой категории только некоторую часть примеров, принадлежащих другим категориям.

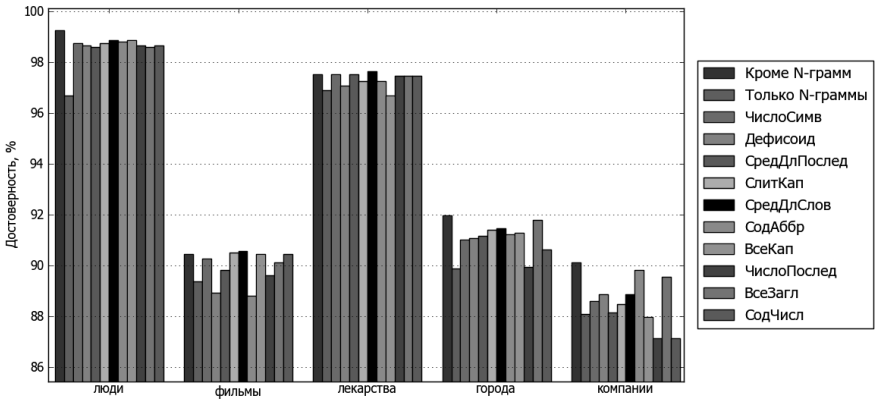


Рис. 4. Классифицирующие комбинации n-грамм и одного дополнительного признака, показатель достоверности, задача выделения одной категории

Для оценки источников ошибок при классификации можно использовать ручной анализ типичных ошибок. Некоторые ошибки приведены в таблице 4.

Таблица 4. Примеры ошибок классификации

Истинная категория	**Результат классификатора**	**Примеры**
Компания	Фильм	Бархатные ручки; Вкус лета; Высота; Гармония; Независимость; Пятерочка; Нокиан Шина; Дикси групп
Компания	Город	Дары Покрова; Волжанин; Корнет; Кронверк; Магнит; Молот; Русал; Север; Останкинское; Кукуй
Город	Компания	Владивосток; Москва; Санкт-Петербург; Шишкин Лес; Янтарный; Йошкар-Ола; Сыктывкар
Город	Фильм	Анна; Свобода; Опочка; Воскресенск; Казахский аул; Прислониха
Лекарство	Компания	АДС; Муромонаб-CD3; Антитромбин III
Фильм	Город	Ад; Бог; Вздох; Дзифт; Дитя; Интервью; Окись; Сын; Халк; Кислород
Фильм	Компания	Generation П; Вдох-выдох; Королева; Консервы; Здравозахоронение; Отважная; Рестлер; Макс Пэйн; Корпорация монстров; Рожденные в СССР

7.5. Выводы из экспериментальных результатов

Анализ ошибок в экспериментах показывает следующие основные причины ошибок.

Наиболее достоверно классифицируемыми категориями являются имена людей и названия лекарств. Это объясняется тем фактом, что данные категории обладают наиболее жесткой структурой, характерными префиксами или суффиксами.

Путём исследования ошибок классификации можно сделать выводы о наиболее статистически значимых характеристиках имён тех или иных категорий, которые приведены в таблице 5.

Таблица 5. Значимые характеристики, полученные в результате ручного анализа ошибок классификации

Категория	Значимые характеристики
Имена людей	Структура, префиксы и суффиксы, число слов
Лекарства	Префиксы, суффиксы, n-граммы, число слов
Компании	Специфические свойства структуры: аббревиатуры, капитализация, дефисы, числа; использование географических названий, написание заглавными буквами.
Город	Часто — короткие, мало других признаков.
Фильм	Использование общей лексики, число слов, наличие некапитализованных слов

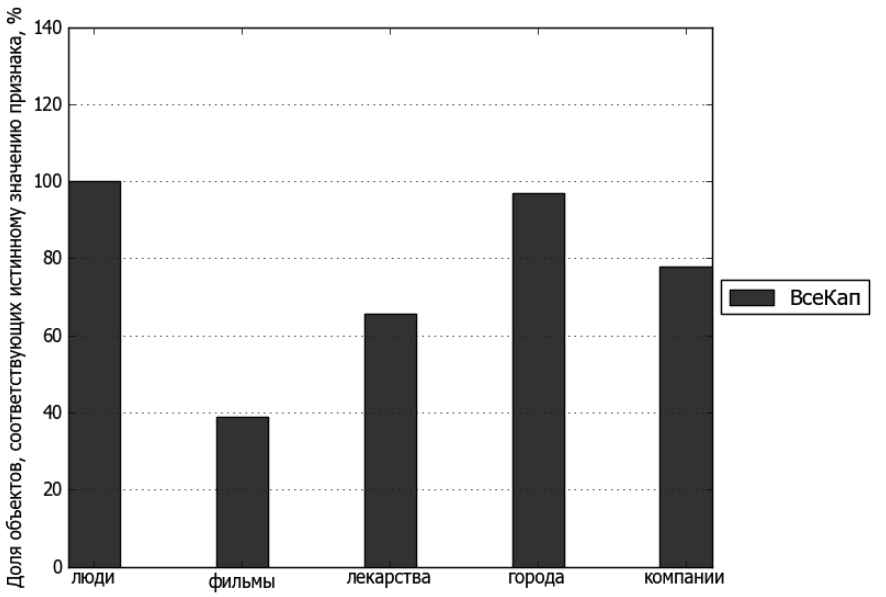


Рис. 5. Доля примеров, в которых все слова капитализованы

Приведём распределения основных признаков по словам из использованного набора. Эти распределения позволяют оценить особенности примеров. По соображениям объёма, они приводятся без дополнительного расширенного описания.

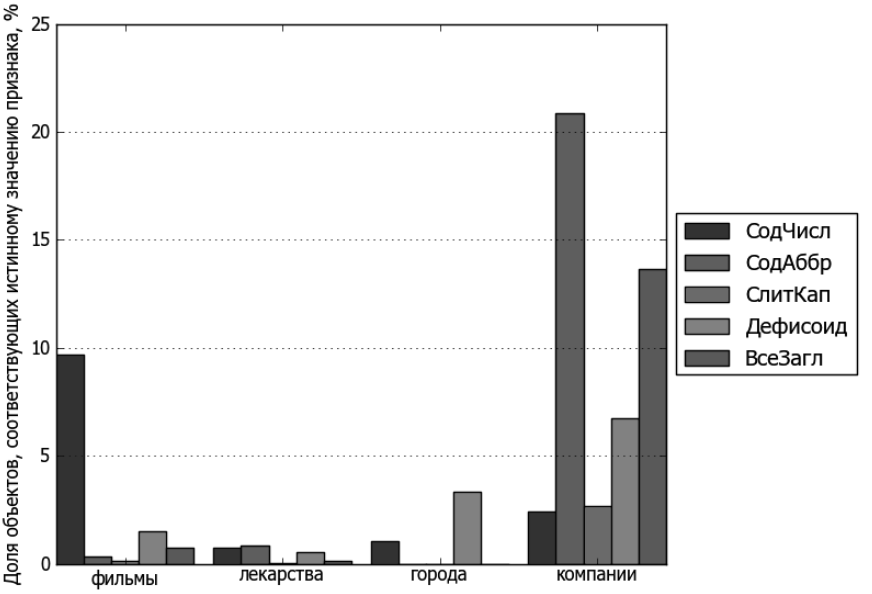


Рис. 6. Распределение истинных значений прочих бинарных признаков по категориям (в категории имён людей все эти признаки имеют значение «ложь»)

8. Сравнение результатов (Patel & Smarr, 2001)

Таблицы 1, 2, 3 показывают, что наиболее значительные отличия в достоверности между нашим классификатором и (Patel & Smarr, 2001) наблюдаются в категориях «Названия городов» и «Названия компаний», где достоверность нашего классификатора значительно ниже. В остальных категориях достигнуты сходные результаты.

Эти отличия объясняются тем, что в нашем наборе названий компаний практически отсутствуют обозначения организационно-правовых форм (*ООО, ОАО, Группа компаний*) и видов деятельности (*Банк*). В то же время в примерах (Patel & Smarr, 2001) эти атрибуты обнаруживаются столь часто (*Inc., INC, Corporation, Fund*), что в цитируемой работе был введён дополнительный внутренний признак — последнее слово имени, обозначаемый в оригинале как «suffix». В нашем случае подобный признак не вносит ощутимого увеличения достоверности классификатора.

В то же время, указанные категории названий городов и компаний обнаруживают существенное сходство в статистике n-грамм, вследствие чего значительная доля ошибок классификации проявляется в неспособности различать именно эти категории.

Заключение

Нами была продемонстрирована возможность создания достаточно простого классификатора, достигающего достоверности классификации 80–99,9 % между двумя категориями, 89–99 % при отделении одной категории от остальных и общей достоверности 85 % при определении категории имени из 5 заранее заданных категорий.

При этом использовались исключительно внутренние признаки, извлекаемые непосредственно из классифицируемых имен и набора обучающих примеров, имеющего объём порядка 2000 примеров для каждой категории.

Полученные результаты аналогичны приводимым в сходной работе (Patel & Smarr, 2001), авторы которой применяли аналогичный подход, но несколько иной набор признаков, для классификации англоязычных имен.

Наиболее перспективным подходом к дальнейшему совершенствованию классификатора, ограниченного использованием сугубо внутренних признаков, представляется использование частотного общелексического словаря для повышения достоверности классификации названий компаний и фильмов.

В наши дальнейшие планы входят испытания классификатора на большем числе категорий, увеличение набора обучающих примеров, поиск дополнительных внутренних признаков. Помимо этого, значительный интерес представляет расширение классификатора путём интеграции его с системой автоматического разбора текста и анализа синтактико-семантической структуры контекста собственного имени — фактически, источника внешней информации.

Литература

1. *Baluja, S.* (2000). Applying Machine Learning for High Performance Named Entity Extraction.
2. *McDonald, D. D.* (1993). “Internal and External Evidence in the Identification and Semantic Categorization of Proper Names.” available at: Acquisition of Lexical Knowledge: http://aclweb.org/anthology-new/W/W93/W93-0104.pdf
3. *Mikheev, A.* (1997). Automatic Rule Induction for Unknown Word Guessing. Computational Linguistics , 23(3), 405–423.
4. *Patel, S., & Smarr, J.* (2001). “Automatic Classification of Previously Unseen Proper Noun Phrases into Semantic Categories using an N-Gram Letter Model.” available at: http://nlp.stanford.edu/courses/cs224n/2001/jsmarr/NGramWordClassifier.pdf

COMBINING MULTIPLE FEATURES FOR SINGLE-WORD TERM EXTRACTION

Nokel M. A. (mnokel@gmail.com)
Moscow State University, Moscow, Russia

Bolshakova E. I. (eibolshakova@gmail.com)
Moscow State University, Moscow, Russia

Loukachevitch N. V. (louk_nat@mail.ru)
Research Computing Center, Moscow State University,
Moscow, Russia

The paper describes experiments on automatic single-word term extraction based on combining various features of words, mainly linguistic and statistical, by machine learning methods. Since single-word terms are much more difficult to recognize than multi-word terms, a broad range of word features was taken into account, among them are widely-known measures (such as TF-IDF), some novel features, as well as proposed modifications of features usually applied for multi-word term extraction.
A large target collection of Russian texts in the domain of banking was taken for experiments. Average Precision was chosen to evaluate the results of term extraction, along with the manually created thesaurus of terminology on banking activity that was used to approve extracted terms.
The experiments showed that the use of multiple features significantly improves the results of automatic extraction of domain-specific terms. It was proved that logistic regression is the best machine learning method for single-word term extraction; the subset of word features significant for term extraction was also revealed.

Key words: single-word, term extraction, combining features, machine learning

1. Introduction

Term extraction is a field of language technology that involves extraction of relevant terms from domain-specific language corpora. To date, the research in this field has tended to focus on extraction of multi-word terms rather than on single-word ones mainly because most terms are multi-word. At the same time it is argued that single-word terms are much more difficult to recognize (Sclano & Velardi, 2007).

An important current trend in the research consists in applying machine learning methods that combine various features of words for term extraction.

In the work (Vivaldi et al., 2001) for extraction of medical terms features of words are combined with AdaBoost algorithm. (Azé et al., 2005) combines 13 various statistical criteria measures via the supervised learning genetic algorithm ROGER.

In (Pecina and Schlesinger, 2006) the combination of multiple statistical characteristics of phrases is used to extract multi-word expressions from the Czech text collection. (Zhang et al., 2008) propose a combined method based on five term recognition algorithms that are capable of handling both single-word and multi-word terms. (Foo and Markel, 2010) apply the rule induction learning system Ripper to automatic term extraction from Swedish patent texts. (Dobrov and Loukachevitch, 2011) combine multiple features for two-word term extraction from texts of two different domains: the broad domain of natural science and technologies and the domain of banking.

Our study continues the described works aiming to apply machine learning in order to improve automatic term extraction, but in contrast to them we deal with single-word terms.

The overall process of extracting single-word terms consists of the following steps:

1) Extraction of term candidates from domain-oriented texts. In our study we consider only nouns and adjectives because they cover the majority of the terms; for our machine learning experiments they are extracted from a target collection of Russian banking texts taken from various electronic magazines such as Analytical Banking Magazine and Auditor.
2) Reordering the list of extracted candidates with the purpose to get more approved terms in the top of the list. To reorder the list, certain word features that measure "termhood" are used. In our study a variety of features (mainly, linguistic and statistical) is considered.

To evaluate the results of the reordering we need a way to approve terms from the candidate list. For these purposes we use the banking thesaurus manually created for the Central Bank of the Russian Federation. It includes about 15 thousand terms and comprises the terminology of banking activity. We consider a given candidate from the list as a term if it belongs to the thesaurus.

In the paper we first characterize the set of chosen features, and then describe experiments with machine learning.

2. Features for term candidate ranking

2.1. Linguistic features

We improve the results of the first step in the term extraction process via the following simple post morphology techniques:

1) We apply a simple morphological disambiguation procedure to extract only those initial forms of the nouns and adjectives that are consistent in text with other context words. Thus, the combinations such as *Preposition + Noun* and *Preposition + Adjective* should be consistent in the case, while the combinations such as *Adjective + Noun*, *Participle + Noun*, *Possessive Pronoun + Noun*, *Ordinal Number + Noun* should be in agreement with the gender, number and case.

2) Term candidates that had the same initial forms with the words with POS other than nouns and adjectives were excluded from the consideration.

For resulted set of term candidates we propose to apply four linguistic features that do not rank the term candidates and are used for the purposes of machine learning: **Ambiguity**, **Novelty**, **POS** and **Specificity**. The first one determines whether the word has multiple initial forms, the second determines whether the word is known for a morphological analyser, the third determines whether the word is a noun or an adjective, and the last determines whether the word exists in the reference text collection.

We also make an attempt to take into account the subjects in the sentences because they are more likely to represent some domain-specific information. All words in the nominative case (according to the morpohological analyzer) are considered as the subjects.

2.2. Orthographic features

Supporting the proposal of (Conrado et al., 2011) we consider the number of occurrences of words beginning with the capital letters. However, we also consider the subset of these words that did not start the sentences because such words are more likely to represent the named entities in the subject domain (they are called non-initial words further in the paper).

2.3. Statistical features

The most of the features of our set are statistical. They may be divided into four groups:

1) Features based only on the target corpus;
2) Features based on the target and reference corpora;
3) Features based on the statistical and contextual information;
4) Features for the words that stand near the most frequent ones in texts.

2.3.1. Used notations

While describing the statistical features, we use the following notations:

– w is the word from the target corpus;
– $TF_t(w)$ and $TF_r(w)$ are the frequencies of the word w in the target and reference corpora;
– $|W_t|$ and $|W_r|$ are the total numbers of words in the target and reference corpora;
– $DF_t(w)$ and $DF_r(w)$ are the numbers of documents containing the word w in the target and reference corpora;
– $|D_t|$ and $|D_r|$ are the total numbers of documents in the target and reference corpora;

2.3.2. Features based only on the target corpus

The most basic features in the group are **Term Frequency** (TF) and **Document Frequency** (DF). The former is the number of occurrences of term candidates in the target corpus, while the latter is the number of documents where a term candidate occurs. These features reflect the assumption that the terms are much more frequent than other words in the target corpus.

The more complex feature is **Term Frequency — Inverse Document Frequency** (TF-IDF) that was originated and is widely used in Information Retrieval. This measure encourages words that occur many times within a small number of documents. Primarily, this feature (Manning and Schutze, 1999) was calculated via the general collection. Later, it was adapted to use only the target corpus. Therefore, we consider these two versions of TF-IDF measure:

$$TF\text{-}IDF^{reference}(w) = TF_t(w) \times \log \frac{|D_r|}{DF_r(w)}; \quad TF\text{-}IDF(w) = TF_t(w) \times \log \frac{|D_t|}{DF_t(w)}$$

We also use an important extension of TF-IDF measure called **Term Frequency — Residual Inverse Document Frequency** (TF-RIDF) proposed by (Church and Gale, 1995):

$$TF\text{-}RIDF(w) = TF_t(w) \times \left(\log \frac{|D_t|}{DF_t(w)} - \left(-\log\left(1 - e^{-\frac{TF_t(w)}{|D_t|}}\right)\right)\right)$$

TF-RIDF is based on the observation that the Poisson model can only fairly predict the distribution of non-content words. Therefore, the deviation from Poisson can be used to predict term informativeness.

The last feature in the group is **Domain Consensus** (DC) (Navigli and Velardi, 2002). This measure simulates the consensus that a term must gain in a community before considered a relevant domain term. It is an entropy-related feature:

$$DC(w) = -\sum_{d_k \in D_t} \left(freq(w, d_k) \times \log\left(freq(w, d_k)\right)\right)$$

where d_k is a document from the target corpus D_t and $freq(w, d_k)$ is the frequency of the word w in a document d_k divided by the total number of the words in d_k.

All these features were calculated four times: for all term candidates, only for subjects, only for words beginning with capital words and only for non-initial words.

2.3.3. Features based on the target and reference corpora

The first feature in this group is **Weirdness** (Ahmad et al., 2007). It compares term frequencies in the target and reference corpora and reflects the basic idea for all measures in the group that these frequencies significantly differ:

$$Weirdness(w) = \frac{TF_t(w)}{|W_t|} \Big/ \frac{TF_r(w)}{|W_r|}$$

Relevance feature (Peñas et al., 2007) is based on the similar idea:

$$\mathrm{Re}levance(w) = 1 - \frac{1}{\log_2\left(2 + \frac{TF_t(w) \times DF_t(w)}{TF_r(w)}\right)}$$

This weight is high for high frequent terms in the target corpus, unless they are also frequent in the reference corpus or appear in a very small number of documents in the target corpus.

Next, we consider several extensions of TF-IDF measure. **Contrastive Weight** (CW) was proposed in (Basili et al., 2001) as a more accurate weight that reflects the specificity of terms with respect to the target domain. It is based on the heuristic that general words should spread similarly across different domain corpus:

$$CW(w) = \log\left(TF_t(w)\right) \times \left(\log \frac{|W_t| + |W_r|}{TF_t(w) + TF_r(w)}\right)$$

Next, Domain Tendency (DT) and Domain Prevalence (DP), which are slight modifications of Weirdness and CW respectively, contribute to the weight, known as **Discriminative Weight** (DW) (Wong et al., 2007). A term that appears frequently in the target corpus will have a low overall DW if it is more specific in the reference corpus:

$$DW(w) = DP(w) \times DT(w), \text{ where } DT(w) = \log_2\left(\frac{TF_t(w)+1}{TF_r(w)+1} + 1\right)$$

$$DP(w) = \log_{10}\left(TF_t(w) + 10\right) \times \log_{10}\left(\frac{|W_t| + |W_r|}{TF_t(w) + TF_r(w)} + 10\right)$$

One more extension of TF-IDF measure is **KF-IDF** (Kurz and Xu, 2002). This feature considers a simple term candidate as relevant if it appears more often than other candidates in the target domain, but occasionally in the reference domain. This weight is defined as follows:

$$KF\text{-}IDF(w) = DF_t(w) \times \log\left(\frac{2}{|D_w|} + 1\right)$$

where $|D_w|$=2, if the word w exists in the reference collection, and $|D_w|$=1 otherwise.

The last feature in the group is **Loglikelihood** that was originally designed for multi-word term extraction and then adapted by (Gelbukh et al., 2010) to single-word term extraction. Since only term candidates whose relative frequency is greater in the

target corpus than in the reference one are taken into account: $\frac{TF_t(w)}{|W_t|} > \frac{TF_r(w)}{|W_r|}$, Log-likelihood is defined as follows:

$$Loglikelihood(w) = 2 \times \left(TF_t(w) \times \log\left(\frac{TF_t(w)}{TF_t^{\exp ected}(w)} \right) + TF_r(w) \times \log\left(\frac{TF_r(w)}{TF_r^{\exp ected}(w)} \right) \right)$$

$$\text{where } TF_t^{\exp ected}(w) = |W_t| \times \frac{TF_t(w) + TF_r(w)}{|W_t| + |W_r|}; \quad TF_r^{\exp ected}(w) = |W_r| \times \frac{TF_t(w) + TF_r(w)}{|W_t| + |W_r|}$$

2.3.4. Features based on the statistical and contextual information

First of all, C-value and its modifications and improvements are included into the group. Originally, **C-value** was proposed to extract multi-word terms (Ananiadou, 1994), but we use its modified version adapted by (Nakagawa and Mori, 2002) for single-word term extraction:

$$MC\text{-}value(w) = TF_t(w) - \frac{\sum_{p \in P_w} TF_t(p)}{|P_w|}$$

where $|P_w|$ is the set of all phrases in the text collection that contain the word w and $|P_w|$ is its cardinality.

The most widely known modification of C-value is **NC-value** (Frantzi and Ananiadou, 1997). This weight incorporates contextual information into C-value for the extraction of multi-word terms and counts how independently the given multi-word term is used in the target corpus. We adapt NC-value to single-word term extraction as follows:

$$NC\text{-}value(w) = \frac{1}{|W_t|} \times MC\text{-}value(w) \times cweight(w)$$

$$\text{where } cweight(w) = \sum_{c \in C_w} weight(c) + 1; \quad weight(c) = \frac{1}{2}\left(\frac{|W_c|}{|W_t|} + \frac{\sum_{e \in W_c} freq(e)}{TF_t(c)} \right)$$

where C_w is the set of context words of the word w, W_c is the set of the term candidates that have c as a context word, $\sum_{e \in W_c} freq(e)$ is the sum of the frequencies of the term candidates that appear with the word c.

We also consider another form of the original NC-value proposed by (Frantzi and Ananiadou, 1999) and modify it to single-word term extraction:

$$MNC\text{-}value(w) + 0.8 \times MC\text{-}value(w) + 0.2 \times CF(w)$$

where $CF(w) = \sum_{c \in C_w} freq(c)$ is the context factor for the word w, C_w is the set of context words of the word w, $freq(c)$ is the frequency of the term candidate c as the context word of the word w.

Next feature in the group is LR (Nakagawa and Mori, 2003) that is based on the intuition that some words are used as term units more frequently than others. It was originally proposed for multi-word term extraction, but we adapt it for single-word term extraction by simply replacing the term units in its definition by context words. We consider two versions of this score: **Token-LR** and **Type-LR**:

$$Token\text{-}LR(w) = \sqrt{l_{token}(w) \times r_{token}(w)}; \quad Type\text{-}LR(w) = \sqrt{l_{type}(w) \times r_{type}(w)}$$

where the left score $l_{token}(w)$ of the word w is defined as the sum of the frequencies of the context words appearing to the left of the word w, the left score $l_{type}(w)$ is the cardinality of the set of such context words and the right scores $r_{token}(w)$ and $r_{type}(w)$ are defined in the same manner.

Since all variants of LR method reflect the numbers of occurrences of the context words, but do not reflect the terms themselves, we also choose FLR method intended to overcome this shortcoming (Nakagawa and Mori, 2003). Similar to LR we consider two variants of FLR score: namely, **Type-FLR** and **Token-FLR**:

$$Token\text{-}FLR(w) = TF_t(w) \times Token\text{-}LR(w); \; Type\text{-}FLR(w) = TF_t(w) \times Type\text{-}LR(w)$$

Additionally we consider several features reflecting the usage of the word in a set of phrases. The first one is **Insideness** (Dobrov and Loukachevitch, 2011). It checks whether the word is used in the same phrase and is intended to reveal truncated word sequences that are the parts of the real terms (note, that the similar phenomenon is modelled by previously described C-value feature). Insideness is defined as follows:

$$Insideness(w) = \frac{F_{\max}}{TF_t(w)}$$

where $F_{\max}$ is the frequency of the most frequent phrase containing the word w.

Another feature is SumN proposed by (Loukachevitch and Logachev, 2010), where N is the number of the most frequent phrases containing the considered term candidate. The feature checks productivity of the word for the formation of domain word combinations. We also modify it for single-word term extraction by excluding term frequency from the denominator:

$$SumN(w) = \frac{\sum_{p \in P_w^N} TF_t(p)}{N}$$

where P_w^N is the set of the N most frequent phrases containing w. We consider **Sum3**, **Sum10** and **Sum50** features.

2.3.5. Features for the words that stand near the most frequent ones in texts

At last, we hypothesize that the terms are more likely to co-occur with the most frequent ones and introduce the novel feature **NearTermsFreq** defined as the number

of the word occurrences in the context window of the several predefined most frequent words. In fact, as such words we take the first ten elements at the top of the term candidate list ordered by TF-RIDF because our experiments showed that it is the best single feature — cf. Table 1). Additionally we apply the original TF-IDF measure, calculated via the general text collection, to NearTermsFreq, thus obtaining the following feature:

$$NearTermsFreq\text{-}IDF_{ref}(\mathrm{w}) = NearTermsFreq(w) \times \log\left(\frac{|D_r|}{DF_r(w)}\right)$$

3. Experiments

3.1. Experimental setup

For experiments we used a target text corpus in the banking domain with 10 422 documents (nearly 15,5 million words) and word frequencies from the reference, more general collection. All described features were calculated for five thousand of the most frequent single-word term candidates extracted from the target collection.

In order to obtain the best combination of the features, we used machine learning methods provided by the freely-available library Weka (*http://www.cs.waikato.ac.nz/ml/weka/*). We performed four-fold cross-validation, which means that every time the training set was three-quarters of the whole list while the testing set was the remaining part.

Among various methods of evaluation we chose Average Precision (Manning and Schutze, 1999) because this measure allows us to evaluate the quality of the term extraction using a single numerical value. Average Precision of a set of all term candidates D with $D_q \subseteq D$ as a set of approved ones among them is defined as follows:

$$AvP(D) = \frac{1}{|D_q|} \sum_{1 \le k \le |D|} \left(r_k \times \left(\frac{1}{k} \sum_{1 \le i \le k} r_i \right) \right)$$

where r_i=1 if the i-th term $\in D_q$ and r_i=0 otherwise. The formula reflects the fact that the more terms are concentrated in the top of the list, the higher the measure is.

3.2. Experimental results

In order to find the best combination of the features we tested several machine learning methods. It proved that the maximal value of Average Precision is achieved by logistic regression method. So it was taken for further experiments.

Table 1 shows AvP values for single features and their combination obtained with logistic regression (Ambiguity, Novelty, POS and Specificity features are not

presented in the table because they do not rank the term candidates and are used only in the combination with the other features).

Table 1. Average Precision for single features and logistic regression

Feature	Average Precision
TF	33,91%
DF	28,7%
TF-IDF	37,84%
TF-RIDF	**40,05%**
DC	32,42%
$TF\text{-}IDF_{reference}$	34,56%
$TF_{subjects}$	29,66%
$DF_{subjects}$	27,92%
$TF\text{-}IDF_{subjects}$	30,56%
$TF\text{-}RIDF_{subjects}$	32,61%
$DC_{subjects}$	28,92%
$TF\text{-}IDF_{reference\ subjects}$	29,61%
$TF_{capital\ words}$	35,49%
$DF_{capital\ words}$	33,42%
$TF\text{-}IDF_{capital\ words}$	35,98%
$TF\text{-}RIDF_{capital\ words}$	37,97%
$DC_{capital\ words}$	34,63%
$TF\text{-}IDF_{reference\ capital\ words}$	35,51%
$TF_{non\text{-}initial\ words}$	36,29%
$DF_{non\text{-}initial\ words}$	36,12%
$TF\text{-}IDF_{non\text{-}initial\ words}$	36,26%
$TF\text{-}RIDF_{non\text{-}initial\ words}$	35,85%
$DC_{non\text{-}initial\ words}$	35,77%
$TF\text{-}IDF_{reference\ non\text{-}initial\ words}$	32,83%

Feature	Average Precision
Weirdness	29,87%
Relevance	32,43%
CW	34,42%
DW	30,37%
KF-IDF	28,68%
Loglikelihood	34,48%
MC-value	33,86%
NC-value	35,1%
MNC-value	34,55%
Token-LR	35,93%
Type-LR	33,21%
Token-FLR	35,44%
Type-FLR	34,02%
Insideness	27,8%
Sum3	36,88%
Sum10	37,22%
Sum50	36,86%
NearTermsFreq	35,76%
$NearTermsFreq\text{-}IDF_{ref}$	36,06%
Logistic Regression	**53,95%**

As we see, the best single feature is TF-RIDF, while logistic regression by combining multiple features gives an increase of 35% compared with the best single feature.

In the Table 2 the first ten elements from the top of the extracted term candidates lists are presented. The columns correspond to various orderings of the lists: the ordering by Term Frequency, by TF-RIDF feature, and by logistic regression (the real terms among them are given in italics).

Table 2. First ten extracted term candidates

#	Term Frequency	TF-RIDF	Logistic Regression
1	*Банк (Bank)*	*Банк (Bank)*	*Банковский (Banking)*
2	*Банковский (Banking)*	*Кредитный (Credit)*	*Компания (Company)*
3	*Россия (Russia)*	*Банковский (Banking)*	*Рынок (Market)*

#	Term Frequency	TF-RIDF	Logistic Regression
4	Год (Year)	Риск (Risk)	Риск (Risk)
5	*Система (System)*	*Кредит (Credit)*	*Пенсионный (Pensionary)*
6	Организация (Organization)	*Рынок (Market)*	*Аудиторский (Auditing)*
7	*Рынок (Market)*	*Система (System)*	*Страна (Country)*
8	*Кредитный (Credit)*	*Налоговый (Taxing)*	*Налоговый (Taxing)*
9	Банка (Jar)	*Страна (Country)*	*Система (System)*
10	*Российский (Russian)*	Банка (Jar)	*Бухгалтерский (Bookkeeping)*

3.3. Feature selection algorithm

The resulting combination model is too complex in the number of applied features. Some of them may be redundant for machine learning method and have no use in the model, make its training harder. In order to exclude them we applied a stepwise greedy algorithm for selecting the most significant features.

The algorithm starts with the empty set of features, and then at each step it adds the feature that maximizes the overall Average Precision. As a result, the combination of only eight features (namely, TF-IDF, TF-RIDF, KF-IDF, $DF_{non\text{-}initial\ words}$, $TF_{subjects}$, $TF\text{-}IDF_{reference\ subjects}$, Weirdness and NC-value) was found with **53,51 %** of Average Precision. Therefore, the number of the combined features may be considerably reduced with decrease in precision less than 1 %.

4. Conclusions

In the paper we described multiple word features including linguistic, orthographic and statistical ones that were used in machine learning experiments for ordering the set of single-word term candidates extracted from the target text corpus. Several machine learning methods combining the features were tested, and logistic regression proved to be the best with significantly higher values of Average Precision than for any single feature. In addition, it was experimentally found that the number of the combined features can be reduced to eight features without sensible decrease of Average Precision.

References

1. *Ahmad, K., Gillam, L., Tostevin L.* University of Survey participation in Trec8: Weirdness indexing for logical document extrapolation and retrieval. Proceedings of the 8th Text Retrieval Conference. Gaithersburg, USA, 2007.
2. *Ananiadou S.* A Methodology for Automatic Term Recognition. Proceedings of the 15th International Conference on Computational Linguistics, COLING'94. Kyoto, Japan, 1994, pp. 1034–1038.

3. *Azé, J., Roche, M., Kodratoff, Y., Sebag, M.* Preference Learning in Terminology Extraction: A ROC-based Approach. Proceedings of ASMDA'05 (Applied Stochastic Models and Data Analysis). Brest, France, 2005, pp. 209–219.
4. *Basili, R., Moschitti, A., Pazienza, M., Zanzotto, F.* A Contrastive Approach to Term Extraction. Proceedings of the 4th Terminology and Artificial Intelligence Conference (TIA). Nancy, France, 2001.
5. *Church, K., Gale, W.* Inverse Document Frequency IDF: A Measure of Deviation from Poisson. Proceedings of the Third Workshop on Very Large Corpora. Cambridge, USA, 1995, pp. 121–130.
6. *Conrado, M., Koza W., Díaz-Labrador, J., Abaitua, J., Rezende, S., Pardo, T., Solana, Z.* Experiments on Term Extraction Using Noun Phrase Subclassifications. Proceedings of the 8th Recent Advances in Natural Language Processing Conference (RANLP 2011). Hissar, Bulgaria, 2011, pp. 746–751.
7. *Dobrov, B., Loukachevitch, N.* 2011. Multiple Evidence for Term Extraction in Broad Domains. Proceedings of the 8th Recent Advances in Natural Language Processing Conference (RANLP 2011). Hissar, Bulgaria, 2011, pp. 710–715.
8. *Foo, J., Merkel M.* Using Machine Learning to Perform Automatic Term Recognition. Proceedings of the LREC2010 Aquisition Workshop. Malta, 2010.
9. *Frantzi, K., Ananiadou, S.* Automatic Term Recognition Using Contextual Cues. Proceedings of the IJCAI Workshop on Multilinguality in Software Industry: the AI Contribution. Nagoya, Japan, 1997.
10. *Frantzi, K., Ananiadou, S.* (1999), The C-value/NC-value Domain-independent Method for Multi-word Term Extraction. Journal of Natural Language Processing, Vol. 6, no. 3, pp. 145–179.
11. *Gelbukh, A., Sidorov, G., Lavin-Villa, E., Chanona-Hernandez, L.* Automatic Term Extraction using Log-likelihood based Comparison with General Reference Corpus. Proceedings of the Natural Language Processing and Information Systems, and the 15th International Conference on Applications of Natural Language to Information Systems. Cardiff, UK, 2010, pp. 248–255.
12. *Kurz, D., Xu, F.* Text Mining for the Extraction of Domain Relevant Terms and Term Collocations. Proceedings of the International Workshop on Computational Approaches to Collocations. Vienna, 2002.
13. *Loukachevitch, N., Logachev, IU.* Using Machine Learning for Term Extraction [Ispol'zovanie metodov mashinnogo obucheniia dlia izvlecheniia slov-terminov]. Trudy Dvenadtsatoi Natsional'noi Konferentcii po Iskusstvennomy Intellekty c Mezhdunarodnym Uchastiem: KII 2010 (Proceedings of the 12th National Conference on Artificial Intelligence with International Participation). Tver', Russia, 2010.
14. *Manning, C. D., Schutze, H.* (1999), Foundations of Statistical Language Processing. The MIT Press.
15. *Nakagawa, H., Mori, T.* A Simple but Powerful Automatic Term Extraction Method. COMPUTERM 2002 — Proceedings of the 2nd International Workshop on Computational Terminology. Taipel, Taiwan, 2002, pp. 29–35.
16. *Nakagawa, H., Mori, T.* (2003), Automatic Term Recognition based on Statistics of Compound Nouns and their Components. Terminology, Vol. 9, no.2, pp. 201–219.

17. *Navigli, R., Velardi, P.* Semantic Interpretation of Terminological Strings. Proceedings of the 6th International Conference on Terminology and Knowledge Engineering (TKE 2002). Nancy, France, 2002, pp. 95–100.
18. *Pecina, P., Schlesinger, P.* Combining Association Measures for Collocation Extraction. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. Sydney, Australia, 2006, pp. 651–658.
19. *Sclano, F., Velardi P.* TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities. Proceedings of the 9th Conference on Terminology and Artificial Language TIA 2007. Sophia Antipolis, 2007.
20. *Vivaldi, J., Màrquez, L., Rodríguez, H.* Improving Term Extraction by System Combination Using Boosting. Proceedings of the 12th European Conference on Machine Learning (ECML 2001). Freiburg, Germany, 2001, pp. 515–526.
21. *Wong, W., Liu, W., Bennamoun, M.* Determining Termhood for Learning Domain Ontologies using Domain Prevalence and Tendency. Proceedings of the 6th Aistralasian Conference on Data Mining (AusDM). Gold Coast, 2007, pp. 47–54.
22. *Zhang, Z., Iria, J., Brewster, C., Ciravegna, F.* A Comparative Evaluation of Term Recognition Algorithms. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). Marrakech, Morocoo, 2008.

БАШКИРСКИЙ ИНТЕРНЕТ: ЛЕКСИКА И ПРАГМАТИКА В КОЛИЧЕСТВЕННОМ АСПЕКТЕ

Орехов Б. В. (nevmenandr@gmail.com)
Российская академия народного хозяйства
и государственной службы при Президенте РФ,
Башкирский государственный университет

Галлямов А. А. (azamat.gallyamov@gmail.com)
Башкирский государственный университет

В работе представляются результаты подсчётов, проведённых в башкироязычном сегменте Интернета. Лексикостатистика показывает ориентированность башкирского интернета не на общение или передачу актуальной информации, а на представление официально-деловых документов и прочие формы представительского присутствия башкирского языка в Сети.

Ключевые слова: башкирский язык, лексикостатистика, прагматика, Интернет

По аналогии с Рунетом постепенно появляются и другие понятия, распространяющиеся на сетевые сообщества, уже не связанные напрямую с доменами первого уровня. Например, вполне устойчивым и употребительным понятием является «Татнет», то есть татарский сегмент всемирной паутины[1]. Существуют национальные сегменты Сети и в зарубежном интернете. Об их объёмах можно судить по размерам Википедии на соответствующем языке. К примеру, на странице статистики интернет-энциклопедии на настоящий момент можно получить такие цифры: каталанский — 370.279 статей, валлийский — 35.614, бенгальский — 23.211, шотландский (гэльский) — 9.792, нижнесаксонский — 4.741. Для сравнения для энциклопедий на некоторых российских национальных языках эти цифры выглядят так: якутский — 8.132, удмуртский — 3.066, кабардино-черкесский — 649. Для башкирского этот показатель составляет 15.935.

Ввиду малого объёма и числа пользователей башкирским интернетом до сих пор не интересовались. Мы попытаемся дать характеристику Башнета на основе того материала, который получен нами в результате нескольких измерительных операций, осуществлённых автоматическим способом.

[1] Существует и статья «Татнет» в Википедии, и специально посвящённая предмету вышедшая двумя изданиями книга Айнура Сибгатуллина «Татарский интернет» [Татнет].

Под Башнетом мы будем понимать совокупность доступных в сети Интернет текстовых документов, написанных на башкирском языке.

По нашим данным на конец января 2012 года объём башкирского сегмента составляет 66.199 страниц[2]. Примечательно, что эта величина не находится в линейной зависимости от численности говорящих на башкирском языке. Достаточно сравнить полученные данные с объёмом Рунета. Постоянно растущий Рунет составлял на декабрь 2005 года приблизительно $2538 \cdot 10^6$ документов [Сегалович и др. 2006], на осень 2009 года $3825 \cdot 10^6$ страниц [Контент]. Таким образом, по самой грубой оценке, современный объём Рунета не может быть менее 5 миллиардов документов, то есть разница с объёмом Башнета составляет 5 порядков. По оценке Росстата население России на 1 января 2012 года составляет 143.030.106 человек [Росстат]. При этом число говорящих на башкирском языке в Российской Федерации оценивалось в 2002 году в 1.380.000 (по данным переписи) [Ethnologue]. Таким образом, разница в числе говорящих составляет два порядка.

Разумеется, Интернет — это особая область проявления языковой компетенции. Помимо числа говорящих для количественной оценки этой сферы имеет значение и число пользователей Интернета среди носителей языка. Президент Башкортостана Рустэм Хамитов в своём блоге в записи от 30 сентября 2011 года назвал следующие цифры: «В республике только чуть более 30% населения постоянно пользуются интернетом. В России — около 40%»[3]. Итак, доля пользователей среди русскоговорящего и башкироговорящего населения сопоставима, а вот пропорции к числу сайтов на этих языках — нет.

Таким образом, несмотря на рост потенциальных пользователей Интернета среди башкир, Башнет остаётся крайне ограниченной областью Всемирной сети.

В проведённом нами замере участвовало 30 доменных имён. Это не выборка, а более-менее полный перечень сайтов, которые вообще содержат тексты на башкирском языке.

Многие из обследованных сайтов являются дву- (и более) язычными. На них присутствуют как тексты на башкирском, так и на русском, и на английском языке. Таким образом, строго говоря, исследовались не все страницы этих 30 сайтов, а только те, которые содержат башкирские слова. Таким образом, нахождение именно башкирского текста на интернет-странице представляет собой некоторую техническую проблему, от решения которой зависят конечные статистические данные. В качестве рабочей гипотезы было принято, что в строке башкирского текста будут обязательно присутствовать слова, содержащие специфические для башкирской графики отсутствующие в русском языке буквы: *ң, ҙ, ә, һ, ө, ҡ, ғ, ү, ҫ*. Только такие строки и брались для расчёта.

Показательно, что в Рунете по данным компании «Яндекс» на 2009 год опубликовано около 2,3 триллиона слов [Контент]. Методика подсчёта

---

[2] Весь объём сайтов, которые публикуют тексты на башкирском языке, составляет 87.462 страницы, из них 21.263 не содержат башкирских текстов.

[3] http://blog-rkhamitov.livejournal.com/44543.html

специалистов «Яндекса» такова, что Башнет входит в эту цифру. По нашим данным в Башнете содержится 27.252.251 слово. Таким образом, разница в объёмах Рунета и Башнета в словоупотреблениях будет составлять те же 5 порядков, что и при сравнении страниц.

Логически сайты Башнета следует разделить на две группы. В первую войдут те сайты, которые не содержат специального веб-ориентированного контента. Это будут страницы республиканских СМИ, которые печатаются на бумаге или выходят в радиоэфир и одновременно с этим вывешивают свою продукцию в Интернете. Мы будем называть такие сайты «медийными». Во вторую группу будут входить все остальные сайты, то есть те, наполнение которых изначально рассчитано на Интернет. Такое деление представляется оправданным, так как современные мыслители считают, что «глобальная сеть формирует новые языковые системы — знаки, символы, гипертексты без конца и начала, “языковые игры”, требующие анализа» [Тарасенко 2000]. Медийные же сайты не вносят в эти новации ничего от себя, так как изначально ориентируются на традиционные формы бытования своей продукции, а сетевое присутствие является для них всего лишь необязательным дополнением, никак не влияющим на содержательный аспект представленных материалов.

Медийная группа сайтов — это примерно треть всего перечня (9 из 30 сайтов). Однако её доля в страницах почти достигает половины (41.298 страниц из 87.462), а в словоупотреблениях даже превышает половину (15.155.408 из 27.252.251 слов). Именно в этой группе находится самый объёмный сайт Башнета — страница газеты «Йәшлек». На нём опубликовано 8.470.444 слова, что составляет 55.89% всего объёма словоупотреблений на медийных сайтах и 31% от всего корпуса Башнета.

Наиболее значительный ресурс в во второй группе — свободная мультиязычная интернет-энциклопедия «Википедия». Блогов, в которых публиковались бы тексты только на башкирском языке, нам обнаружить не удалось, но нашлись два личных дневника, в которых такие тексты появлялись время от времени, перемежаясь с русскими.

Тематический и жанровый обзор свидетельствует, что для башкирского сегмента Сети фактически отсутствуют все основные сервисы, обычно занимающие большую часть времени пользователя. Иными словами, нет того, зачем обычно приходят в Интернет: электронная почта, поисковая система, информация о погоде, современные новостные ресурсы, социальные сети.

Как минимум на треть Башнет является всего лишь приложением к СМИ. В гораздо меньшей степени, чем это развито в Рунете и всемирной паутине в целом, Башнет является ресурсом общения. Это означает, что, по большому счёту, размещённые в Интернете башкирские тексты не предполагают существования читателя. Заходящий в сеть носитель башкирского языка воспользуется для своих повседневных нужд сайтом на другом языке, просто потому что таких востребованных сервисов в Башнете нет. Главным образом, башкирские сайты экстравертны — они обращены к внешнему миру, к сообществу людей, которые не являются носителями башкирского языка, и главным их сообщением является сигнал о существовании башкирского языка как такового.

По частотности лексики язык Интернета заведомо отличается от языка печатных изданий (это ещё одна причина разделять сайты на медийные и остальные). На материале русского языка было сделано наблюдение, что «существительные, распространенные в текстах на сайтах и в письменных бумажных текстах, совпадают очень мало. Это неудивительно: топ-20 популярных в интернете существительных наполовину состоит из интернет-терминов» [Контент]. Для Башнета это не так. Самым частотным словом из лексики сетевой коммуникации в башкирском интернете выступает «форум», который находится в средине третьего десятка. Следующий за ним специфический интернет-термин — «Википедия» — обнаруживается в середине четвёртого десятка.

Очевидно, что такое положение вещей обусловлено неразвитостью интернет-коммуникации на башкирском языке и, соответственно, отсутствием сетевого метаязыкового уровня.

Для лучшего уяснения ситуации с лексическим составом верхней части частотного словаря Башнета можно использовать сравнение получившегося в результате обследования башкирских сайтов частотного списка с существующими частотными словарями башкирского языка. Мы использовали составленные З. А. Сиразитдиновым словари на основе научных текстов (Н), художественной прозы (П) и произведений писателя Даута Юлтыя (Ю) [Сиразитдинов 1997, Сиразитдинов 2002, Частотный].

Соединительный союз *һәм* ‘и’ и послелог *менән* ‘с’ вполне ожидаемо занимают верхние строчки в частотнике Башнета. Столь же высокие позиции у этих слов и в указанных частотных словарях башкирского языка.

Но вот третье слово в частотном словаре Башнета в сравнении с другими частотниками довольно неожиданно, и, как мы считаем, отражает крен в сторону официального языка. Это послелог *буйынса* ‘поэтому, вследствие этого’ (78.230 вхождений), который не встречается в числе 50 наиболее употребляемых слов ни в одном из частотных словарей башкирского языка. Употребительность этого послелога может быть связана с такими случаями, как этот: *баш мөхәррирҙең дөйөм мәсьәләләр буйынса урынбаҫары* ‘заместитель редактора **по** общим вопросам’. Такого рода контекстов достаточно много, что лучше всего иллюстрирует мысль о том, что Башнет в его нынешнем виде, по сути, не предназначен для общения и более всего отражает норму официально-деловых текстов.

Четвёртое место в частотнике занимает слово *башҡорт* ‘башкирский’ (66.945 вхождений). В других словарях: Н: 15 место (1.126 вхождений); в П и Ю: не входит в первые 50 слов (в Ю: 68 место). Как и слово *башҡортостан* (17 место и 43.272 вхождений), это, прежде всего, слово официального языка, а кроме этого ещё и главный элемент в ряду знаков саморепрезентации языка.

Неожиданно высокие позиции в частотном списке занимают формы *ағасым* ‘моё дерево’ (6 место и 59.792 вхождений), *ҡошом* ‘моя птица’ (8 место, 59.739 вхождений), *ырыуым* ‘мой род’ (10 место, 47.789 вхождений), *ораным* ‘мой клич’ (16 место, 43.791 вхождений). Такое необычное положение вещей объясняется форматом башкирского форума (http://www.bashforum.net/), регистрация на котором предполагает указание пользователем признаков

единства своего племени[4], которые отображаются на каждой странице рядом с добавленной им репликой. Таким образом, формы *ағасым*, *ҡошом*, *ырыуым* и *ораным* встречаются на сайте столько раз, сколько реплик пользователей он показывает.

Перенесённая в Сеть традиционная башкирская форма определения себя в сопоставлении с окружающими, особенно знаменательна на фоне высказанной выше гипотезы о самом существовании Башнета как способе определения языка в современных условиях информационного общества.

Глагол *үҙгәртергә* ‘изменить’ (9 место и 49.017 вхождений в частотном списке Башнета) тоже не характерное часто употребляемое слово в башкирском языке за пределами Интернета. Его высокие позиции в перечне лексики объясняются дизайнерским решением «Википедии», в которой слово «изменить» сопровождает каждый раздел статьи (которых на одной странице может быть много). А так как «Википедия» составляет 31,28 % всего объёма немедийного Башнета, частотности слова *үҙгәртергә* удивляться не следует. То же относится и к слову *мәҡәлә* ‘статья’ (12 место 45.984 вхождений), отсутствующему в ряду самой частотной лексики в Н, П и Ю. При этом можно было бы ожидать высокие позиции этого слова в Н, однако, в реальности мы сталкиваемся с иным положением вещей.

Специального очерка заслуживает слово *йүнәлештәр* ‘направления’ (14 место, 44.049 вхождений). Его мы тоже вряд ли смогли бы ожидать на верхних позициях частотного списка, составленного на основе сбалансированного корпуса текстов данного языка (в Н, П и Ю не входит в число наиболее частотных). Однако чрезмерно канцеляризированный язык Башнета демонстрирует аномальную популярность этой словоформы. Она появляется в характерных контекстах, в которых речь идёт о развитии новых направлений, стремлении к новым направлениям исследований, о работе по следующим направлениям, об основных направлениях работы. Самая частая коллокация с участием этой лексемы в корпусе Башнета: *төп йүнәлеш* ‘основное направление’. Обилие контекстов свидетельствует, что *йүнәлеш* относится к ряду особенно «модных» слов с предельно абстрактным значением, об обилии которых в современной русской языковой ситуации писал М. А. Кронгауз [Кронгауз 2007] (ср. также *нанотехнологии*, *инновации*, *модернизация* и т. д.). Вот несколько репрезентативных иллюстраций: *ә 2009 йылдан факультетта бөтә йүнәлештәр буйынса ла студенттарҙың олимпиадаһы уҙғарыла* ‘а с 2009 года на факультете проводятся олимпиады по всем направлениям’; *Баймаҡ районында тап ошо йүнәлештәрҙе үҫтереү өсөн бөтә мөмкинлектәр бар* ‘в Баймакском районе для развития вот этих вот направлений есть все возможности’.

Ещё одну неожиданность составляет 15 позиция и 43.955 вхождений словоформы *бейҙәр* ‘бии, князья’, отсутствующей среди частотных в Н, П и Ю. Высокую частотность этому слову обеспечивает «Википедия».

[4] См. статью «Структура родоплеменной организации башкир» в Краткой энциклопедии «Башкортостан» (1996).

Что характерно, мы не находим слов *буйынса*, *ағасым*, *ҡошом*, *ырыуым*, *үҙгәртергә*, *мәҡәлә*, *йүнәлештәр*, *бейҙәр* в верхней части частотного списка, составленного нами на материале сайтов башкирских СМИ. В то же время наиболее закономерно употребительные служебные слова *һәм*, *был*, *менән*, *өсөн* присутствуют и там, и там.

Отдельную проблему составляет форма *йылға* (19 место, 40.237 вхождений), которая одновременно может быть и сущ. в им. п. ‘река’, и сущ. в дат. п. ‘год’. Таким образом, частотность этой формы складывается из всех энциклопедических статей Википедии, в которых так или иначе упоминаются реки, и всех контекстов, где этим словом обозначен некоторый промежуток времени, что особенно востребовано на интернет-страницах.

Особый интерес представляют формы, встречающиеся чаще в печатных текстах, чем в Башнете. К таким можно отнести, например, неизменяемую форму *ине*, употребляемую в прошедшем времени со значением усиленной неопределённости: *ҡайтты* ‘он (точно) вернулся’; *ҡайтҡан* ‘он (вроде бы, я не видел) вернулся’; *ҡайтҡан ине* ‘он, кажется, вроде, вернулся’. Для научных текстов такие конструкции нехарактерны, а в П (11 место) и в Ю (16 место), напротив, демонстрируют свою востребованность. По сравнению с этими показателями Башнет демонстрирует свою умеренную заинтересованность в форме *ине* (78 место, 18.393 употреблений), что объясняется малым количеством коммуникативно-ориентированных текстов.

Интересным показателем жанрового состава попавших в базу документов являются обсценные слова. Их наличие свидетельствовало бы о ситуации живого общения. Здесь мы наблюдаем довольно ущербную парадигму, в которой актуализированы лишь слова со значением ‘meretrix’ (*уйнаш*, *уйнашсы*: 21 и 105 употреблений соответственно), а остальные лексемы, связанные с именованием телесного низа и под., в Башнете не встречаются.

Среди слов, имеющих высокую частотность и в Рунете, и в Башнете, можно упомянуть *год / йыл* (3 место в списке существительных Рунета[5] / 50 место для начальной формы в Башнете и сравнительно высокие позиции у омонимичных форм типа *йылға* в общем списке, см. выше[6]), *новость / яңылыҡтар* (4 место в Рунете / 176 место в Башнете, 12.530 вхождений), *форум / форум* (7 место в Рунете / 25 место в Башнете), *поиск / эҙләү* (8 место в Рунете / 26 место в Башнете, 33.217 вхождений), *день / көн* (9 место в Рунете / 339 место, 6.137 вхождений), *пользователь / ҡулланыусы* (14 место в Рунете / 562 место в Башнете, 3.333 вхождений), *время / ваҡыт* (19 место в Рунете / 199 место в Башнете, 10.790 вхождений), *человек / кеше* (20 место в Рунете /

[5] Данные взяты из [Контент].

[6] Цифры для сравнения весьма приблизительные, так как рейтинги для Рунета и Башнета составлены по разным принципам: в Рунете считаются леммы, а в Башнете словоформы; в Рунете каждая часть речи приводится в составе собственного рейтинга, а в Башнете учитывается место слово в общем (смешанном с точки зрения частеречной принадлежности) рейтинге. Тем не менее, есть слова, для которых разница в употребимости достаточно показательна, несмотря на принципы подсчёта.

93 место в Башнете, 16.745 вхождений). В этом списке видны ключевые слова, отражающие, как можно представить, универсальные для сайтостроительства понятия, независимо от языка оккупирующие верхнюю часть списка. Не менее показателен список слов, которые характерны для Рунета, но редко встречаются в Башнете. Он демонстрирует, для чего, в отличие от Рунета, Башнет пока не предназначен: *телефон / телефон* (5 место в Рунете / 2.721 место в Башнете, 474 вхождения) *регистрация / регистрациялау* (16 место в Рунете / 107.110 место в Башнете, 4 вхождения), *комментарий / комментарий* (18 место в Рунете, 11.965 место в Башнете, 75 вхождений) *товар / тауар* (13 место в Рунете / 6.689 место в Башнете, 153 вхождения). Таким образом, Башнет не может быть рабочим инструментом для того, в чём давно преуспел Рунет: в нём нельзя искать телефоны, товары, оставлять комментарии, в нём мало сайтов с возможностью регистрации (а значит, привлекающих постоянную аудиторию пользователей).

В частеречной группе прилагательных особенно репрезентативна частотность адъектива *мобильный* (*мобиль* в башкирском). В Рунете это слово находится на 7 месте в перечне прилагательных, в Башнете его частотность равна 1 употреблению (в нашем списке оно попало на 233.607 позицию).

Итак, как можно заключить из частотного анализа лексики Башнета, его язык демонстрирует закономерное своеобразие по сравнению с языком печатных текстов. Своеобразие это продиктовано не столько специфическими реалиями сетевой коммуникации, сколько дизайнерскими решениями вебстраниц, соединёнными с сильным акцентом на официально-деловой стороне публикуемых документов.

Литература

1. *Контент* Рунета [Электронный ресурс] // URL: http://company.yandex.ru/researches/reports/ya_content_09.xml
2. *Кронгауз М. А.* Русский язык на грани нервного срыва. — М.: Языки славянских культур, 2007.
3. *Ляшевская О. Н., Шаров С. А.* Частотный словарь современного русского языка на материалах Национального корпуса русского языка. — М.: Издательский центр «Азбуковник», 2009.
4. *Росстат*. Демография [Электронный ресурс] // URL: http://www.gks.ru/wps/wcm/connect/rosstat/rosstatsite/main/population/demography/
5. *Сегалович И. В.* Методы сравнительного анализа современных поисковых систем и определения объема Рунета [Электронный ресурс] / И. В. Сегалович, Ю. Г. Зеленков, Д. О. Нагорнов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Восьмая Всероссийская научная конференция. Суздаль, 17–19 октября 2006 года. — Суздаль, 2006. — URL: http://download.yandex.ru/company/paper_76_v1.pdf
6. *Сиразитдинов З. А.* Частотный словарь башкирского языка. Т.1 (наука). — Уфа: Гилем, 1997. — 330 с.

7. *Сиразитдинов З. А.* Частотный словарь башкирского языка. Т.2 <в выходных данных ошибочно: «Т.1»> (проза). — Уфа: Гилем, 2002. — 413 с.
8. *Тарасенко В. В.* Антропология Интернет: самоорганизация человека кликающего // Общественные науки и современность. 2000. № 5
9. *Татнет* // Википедия. [2011–2012]. Дата обновления: 15.11.2011. URL: http://ru.wikipedia.org/?oldid=39263025 (дата обращения: 1.2.2012).
10. *Частотный* словарь языка произведений Даута Юлтыя / Составитель З. А. Сиразитдинов / УНЦ РАН. АН РБ. — Уфа, 1995. — 292 с.
11. *Ethnologue* report for language code: bak [Электронный ресурс] // URL: http://www.ethnologue.com/show_language.asp?code=bak

О СПОСОБАХ ЛЕКСИКОГРАФИЧЕСКОГО ОПИСАНИЯ РУССКОЙ СПОНТАННОЙ РЕЧИ (КОРПУСНОЕ ИССЛЕДОВАНИЕ)

Осьмак Н. А. (natalia.osmak@yandex.ru)

Санкт-Петербургский государственный университет
Санкт-Петербург, Россия

В докладе представлены результаты анализа возможных способов лексикографического описания спонтанной речи в целях создания словаря русской бытовой речи. Материалом для анализа стал блок «Один речевой день» Звукового корпуса русского языка. Работа над словарем русской бытовой разговорной речи может осуществляться через описание функционирования наиболее частотных лексем/словоформ или через лексикографический анализ отдельных лексико-семантических групп. В статье последовательно рассматриваются оба способа, а также анализируются их достоинства и недостатки.

Ключевые слова: номинация женщины, номинация мужчины, корпусное исследование, проблемы лексикографии, спонтанная речь

В настоящее время, наряду с существующей академической лексикографией, ставящей перед собой в качестве цели «отражение только тех фактов лексики и семантики, которые характерны для стандартного современного общелитературного, общеразговорного употребления как в его письменной, так и в устной форме; ничего индивидуального, ничего окказионального» (Герд 1990: 29–30), стала появляться традиция «речевой лексикографии», основные принципы которой до конца еще не разработаны. Так, все острее ощущается необходимость создания специального словаря, представляющего реальную речевую ситуацию, те лексические и грамматические единицы, которые уже давно фиксируются в различных корпусах и базах данных, однако в большинстве словарей до сих пор не находят отражения. Данная задача, в свою очередь, требует обращения непосредственно к устной спонтанной речи, поскольку именно она является наиболее полным и ярким отражением состояния речи на настоящий момент.

Такую возможность предоставляет Звуковой корпус русского языка (ЗКРЯ), особенно та его часть (блок «Один речевой день» — ОРД), которая фиксирует повседневную бытовую речь[1] носителей русского языка в ее наиболее естественной форме. Методика 24-часовой записи, используемая при формировании данного блока ЗКРЯ, обеспечивает почти полную свободу говорящего от диктофона

[1] В настоящий момент материал ОРД представляет собой около 300 тыс. словоупотреблений в расшифровках, около 400 часов звучания, записанных у 50 информантов и 600 коммуникантов.

и от тех или иных коммуникативных заданий (чтение, пересказ, описание изображения, рассказ на заданную тему), которые неизбежно усложняют его задачу и обеспечивают «на выходе» хоть и вполне спонтанную, но все же экспериментальную речь (см. подробнее о Звуковом корпусе: Богданова и др. 2008; Степанова 2008). В пользу обращения к ЗКРЯ говорит и то, что именно корпусный подход «отражает точку зрения многих лингвистов на то, каким должно быть адекватное “эмпирическое” изучение языка» (Meyer 2004: 141), а также позволяет выявить новые тенденции, новые языковые явления, отражающиеся, прежде всего, в индивидуальной речи конкретных носителей языка.

Материал Звукового корпуса может быть использован для различных аспектов лексикографического описания спонтанной речи, в том числе для создания, например, словарей контекстных экспрессем русской разговорной речи, дискурсивных единиц и редуцированных форм (см.: Богданова, Осьмак 2011). Однако наиболее перспективной представляется работа над ***Словарем русской бытовой разговорной речи***, поскольку ни один из существующих словарей не дает возможности в полной мере проследить особенности употребления лексических единиц в современной спонтанной речи носителей языка.

Представляется, что работа по составлению такого словаря может осуществляться различными способами. Во-первых, через описание функционирования наиболее частотных лексем/словоформ, а во-вторых, через лексикографический анализ отдельных лексико-семантических групп. Рассмотрим каждый из этих способов подробнее.

Стоит отметить, что целью настоящей статьи является не проспект нового словаря и не полномасштабный лексический анализ тех или иных единиц или групп единиц, а обсуждение различных способов лексикографического описания русской спонтанной речи.

На первый взгляд, обращение к наиболее частотным речевым единицам кажется самым логичным способом начать лексикографическое описание русской устной спонтанной речи. Это подразумевает выявление основных проблем такого описания, а также конкретных тенденций лексических изменений. Более подробный анализ «верхушки» частного словника, полученного в результате обработки записей корпуса «Один речевой день», позволяет судить о том, насколько данный способ применим в целях составления словаря бытовой разговорной речи.

В таблице ниже представлены 150 словоформ этой «верхушки», ранжированных по частотности.

Таблица 1. Наиболее частотные словоформы русского языка (по материалам корпуса «Один речевой день»)

Ранг	Слово	Кол-во	Ранг	Слово	Кол-во	Ранг	Слово	Кол-во
1	я	5 823	51	за	544	100,5	для	294
2	в	5 598	52	здесь	539	102	от	281
3	не	5 480	53,5	будет	538	103	сегодня	273
4	вот	5 195	53,5	очень	538	104	тогда	267
5	ну	5 146	55	потом	537	105	нам	264

Ранг	Слово	Кол-во	Ранг	Слово	Кол-во	Ранг	Слово	Кол-во
6	да	5096	56	когда	530	106	ж	261
7	а	4867	57	тебе	519	107	быть	255
8	что	4460	58,5	такой	512	108,5	три	254
9	и	4213	58,5	тут	512	108,5	пока	254
10	это	3655	60	говорит	506	110	почему	250
11	то	3386	61	или	499	111	них	248
12	там	3267	62	его	493	112	была	245
13	у	3103	63	может	485	113	думаю	244
14	с	3087	64	значит	484	114	такие	242
15	э	2894	65	чего	483	115	него	241
16	на	2736	66	хорошо	475	116	какие	240
17	так	2671	67	ой	474	117	сколько	239
18	как	2240	68	знаешь	468	118	ему	234
19	ты	1798	69	тебя	458	110	во	227
20	всё	1754	70	только	452	120	этого	225
21	нет	1655	71,5	давай	440	121,5	пять	216
22	он	1611	71,5	говорю	440	121,5	понятно	216
23	мне	1284	73	потому	438	123	типа	212
24	угу	1248	74	можно	429	124	слушай	211
25	она	1212	75	кто	406	125	нормально	210
26	есть	1157	76	ничего	405	126	куда	208
27	меня	1118	77	ага	395	127	эта	207
28	по	1103	78	этот	394	128	общем	204
29	они	1069	79	короче	386	129	время	202
30	мы	1066	80	конечно	377	130	наверное	201
31	сейчас	1058	81,5	такая	364	131,5	могу	196
32	еще	1011	81,5	раз	364	131,5	туда	196
33	уже	956	83	до	353	133	много	192
34	но	935	84	о	350	134,5	ей	191
35	бы	927	85	вас	349	134,5	х…й	191
36	надо	923	86	такое	340	136	самое	190
37	же	893	87,5	какой	335	137,5	понимаешь	189
38	просто	864	87,5	их	335	137,5	нужно	189
39	м	782	89	чтобы	332	139	эту	186
40	вообще	763	90	даже	330	140	этом	184
41	к	701	91	блин	325	141	чем	182
42	знаю	693	92	эти	323	142	какая	181
43	нас	682	93	два	318	143	сюда	180
44	если	661	94,5	её	313	144	двадцать	179
45	вы	639	94,5	из	313	144	со	179
46	тоже	629	96	был	306	144	хочу	179
47	все	609	97	один	302	147	вчера	176
48	бл..ь	599	98	вам	301	148	моему	173
49	где	588	99	ли	300	149	больше	170
50	было	562	100,5	ладно	294	150	четыре	169

Рассмотрим подробнее первые десять позиций этого частотного словника. Как видно из таблицы, самым частотным оказалось местоимение *я* (5 823 употреблений). Второе место занимает предлог *в* (5 598 употреблений). Далее следует отрицательная частица *не* (5 480 употреблений), при этом в некоторых ситуациях *не* является редуцированной формой слова-предложения *нет* и иногда употребляется в составе конструкций ***не**-а* и ***не-не-не***, например:

- ***не** / ну она мне сказала / что да вот / она вот случайно;*
- *а **не-не-не** / нам нужно завернуть направо / у меня уже..;*
- ***не-не-не** вот я уже поняла;*
- *ты читала ? # **не-а**.*

Кроме того, зачастую частица *не* употребляется в составе хезитационных конструкций, таких как ***не** это самое, (я) **не** знаю, (я) (**не**) думаю (что)* и под., которые стоит рассматривать не только в качестве единого целого при лексикографическом описании, но и в составе совсем другого словаря — например, дискурсивных единиц, или хезитационных конструкций (см.: Богданова 2011а, б; 2012; Хан 2012).

Далее в «верхушке» частотного словника следуют частицы (тоже по преимуществу хезитативы) *вот, ну, да* (5 195, 5 146 и 5 096 употреблений соответственно), уже давно привлекающие внимание лингвистов. Обратимся лишь к некоторым из последних работ. Так, например, П. Бокале (Бокале 2011а, б) говорит о преимущественной инициальной и финальной позиции слова *вот*, где оно служит для установления контакта либо, соответственно, маркирует завершение планируемого отрезка речи. Ю. В. Дараган выделяет следующие функции данного слова: 1) концептуализация; 2) контроль; 3) оценка; 4) апелляция (Дараган 2011). Особо хотелось бы отметить работу А. И. Кислощук, в которой выделены такие функции дискурсивного слова *вот* в устной спонтанной речи (на материале сбалансированного блока ЗКРЯ), как указательная и усилительная частицы (как правило, в составе различных конструкций, не зафиксированных в существующих словарях), поиск слова или иного продолжения речи, заполнитель паузы хезитации и дискурсивный маркер (Кислощук 2011).

Частица *ну*, являясь текстовым коннектором, часто употребляется в начале фрагмента текста, что, по мнению О. А. Лаптевой, является «маркированной функцией» (Лаптева 1999: 492). Кроме того, *ну* может использоваться для уточнения сказанного или для побуждения, может стимулировать продолжение прерванного высказывания, выражать отказ от первоначального намерения говорить или выступать в противительно-выделительном значении, заменяя при этом союз *но* (Bocale 2011б).

Частица *да* используется в ОРД в значениях, фигурирующих в толковых словарях:

- ***да** (:) нашли //* — выражение утверждения, согласия;

- *а я был один раз на () маслобойке вот на Украине // **да ?** в Донецке () куда я ездил // мы возили семечки в обмен на масло такое знаешь () ну ароматное //* — выражение вопроса при желании получить подтверждение чему-либо;

- ***да** ладно ! *П чем помогала //* — придание высказыванию большей выразительности.

Кроме этого, *да* встречается в качестве заполнения пауз хезитации, как самостоятельно, так и в составе хезитационных конструкций *да вот*, *ну да*:

- *как приводим (...) к единой шкале / **да** / к стандартизированным баллам / мы открываем с вами страницу / триста восемнадцать*;

- *дальше / вот пока с этим понятно / **да** / как в стены переводить ? # **ну да** // # вам понятно ? @ вам всем там / понятно ?*;

- *у нашего испытуемого / **да вот** / *П мы смотрим средний / **да** ... *П пять баллов / пять стенов / средний уровень (...) невротизации / то есть / по сути дела что мы можем *Н как его охарактеризовать*.

Текстовый коннектор *а*, служащий для оформления вопроса, сцепления реплик и введения новой темы является, по мнению П. Бокале, одним из наиболее часто встречающихся, особенно в стартовой позиции (Бокале 2011б):

- ***а** интересно ...*;

- *так / **а** теперь куда ?*

- ***а** зажигалку дала ?*

- ***а** теперь куда / ты видишь ?*

- *четыре ... / **а** четыре вот здесь*.

Финальное *а,* сигнализирующее о просьбе реакции, отклика со стороны собеседника, встречается реже:

- *почему панели такие узкие / **а** ?*

- *ну мне кажется / мне надо больше # какой нам в коридор лучше / **а?***

Восьмое и девятое места в «верхушке» частотного словника ОРД занимают служебные слова *что* и *и* (4460 и 4213 употреблений соответственно). Наконец, замыкает десятку местоимение *это* (3655 употреблений), которое также

входит в ряд таких хезитационных конструкций, как *не это самое*, *это самое* (см. Богданова 2011б; 2012), да и само может быть хезитативом.

Таким образом, беглый взгляд на «частотную верхушку» ОРД дает понять, что лексикографическое описание по этому пути приведет к описанию не лексического состава речи, а, скорее, только дискурсивных или служебных слов.

Первый полноценный глагол в частотном словнике ОРД — *знаю* (693 употреблений; 42 место) При этом ни одного полноценного существительного в списке 150 самых частых слов не обнаружилось вообще. На 48 месте оказалось слово *бл…ь* (599), сопоставимы с ним по употребительности следующие единицы:

- *Блин* — 325 (91 место);
- *Типа* — 212 (123 место);
- *Время* — 202 (129 место).

При этом *блин* употребляется в ОРД на 100% в роли не существительного, а междометия (просторечного, грубого, бранного):

- *нам надо было* ***блин*** ! *где Мойка* ? *как проще вернуться* ? *здесь* ? *туда* ? *или с этой стороны* ?;
- *вот дело в том что знаешь / это в общем короче да / и муж / вот этой подруги взвился говорит / ты вообще там идиотка* ***блин*** */ и я это всё видел*.

Частеречная принадлежность слова *типа* может существенно различаться в зависимости от контекста. Данная лексема может выступать в качестве и частицы, и союза, и существительного, и даже для заполнения паузы хезитации (Богданова 2010а: 6–8).

Схожая ситуация наблюдается со словоформой *время*, которая, наряду с традиционным употреблением в качестве имени существительного, встречается и в других — адвербиальных — значениях (в составе различных конструкций), среди которых *в ближайшее время* (= совсем скоро), *время от времени* (= иногда), *все время* (= постоянно, не переставая), *на время* (= временно, на какой-то срок) и т. д. (Богданова 2010а: 9–11).

Несмотря на кажущуюся логичность этого пути лексикографического описания устной спонтанной речи — обращения к «верхушке» частотного словника как к основному материалу такого описания — от него пока приходится отказаться. Это объясняется, как уже упоминалось выше, тем, что наиболее частотные лексемы словника ОРД представляют собой, в основном, служебные слова, хезитативы и другие дискурсивные единицы, а значит, лексикографическое описание таких «слов» приведет к созданию не словаря русской разговорной бытовой речи, а, скорее, словаря служебных или дискурсивных единиц. Второе же направление — анализ отдельных лексико-семантических

групп — напротив, ставит во главу угла семантику слова, а не его функции, и позволяет проследить семантические, коннотативные и стилистические изменения в употреблении тех или иных лексем. Работа в этом направлении уже позволила получить ряд интересных результатов.

Материалом для пилотного исследования стали две группы русских слов сравнительно небольшого объема — единицы, обозначающие лиц женского и мужского пола. Критерием отбора материала для анализа было наличие у слова наиболее общего значения[2]:

1) общие наименования (*женщина*, *баба*, *мужчина, мужик, дядя*);

2) возрастные наименования (*девочка*, *бабушка, юноша, старик*);

3) наименования с учетом семейного статуса — замужняя/женатый (*женщина — девушка*, *девка, мужчина — парень*)[3].

Первым этапом исследования стало создание корпуса конкретного материала — методом сплошной выборки из компьютерной версии «Малого академического словаря» (далее — МАС)[4]. В итоговый список вошли 40 лексем для лиц женского пола и 38 — для лиц мужского пола.

Следующим шагом анализа стало обращение к корпусу ОРД. После расшифровки записей в программе e-Kar (электронная картотека) был составлен конкорданс всех интересующих нас единиц. Затем данные были экспортированы в программу eLex, позволяющую составлять словарные статьи, подключая контексты ОРД, а также значения слов по МАС. Интерфейс данной программы выглядит следующим образом (см. рисунок).

Многие единицы из исходного списка в ОРД попросту не встретились, и материал для дальнейшего лексикографического анализа сократился до 16 лексем в первом случае и 12 во втором.

Результаты исследования показали, что при попытке лексикографического описания слов, обозначающих лицо мужского или женского пола, наблюдаются следующие явления:

1) совпадение значений МАС и значений, выделяемых из контекстов ОРД (*мужчина, мальчишка, женщина, девушка*);

2) появление новых лексических единиц данного типа, не зафиксированых в МАСе (*дедун, мужичок, мужчинка, бабец, бабуська, бабулька, молодой человек*);

3) особые случаи сочетаемости лексем (*дед, дядя, тетя*);

4) отсутствие значения по ОРД в МАСе (*мальчик, девочка, девчонка*).

---

[2] За пределами исследования были, соответственно, оставлены многие слова, имеющие схожую семантику, такие, например, как термины родства, наименования лица с учетом внешности, социального положения, особенностей характера и т. п. Представляется, что все эти группы слов, отражающих лингвокультурологические особенности определенной эпохи, требуют отдельного изучения.

[3] В данной работе не рассматривались также слова, которые могут быть использованы для обозначения лица как мужского, так и женского пола (ребенок, ребята, ребятишки, дитя и пр.), а также все единицы — обращения к лицу.

[4] Евгеньева А. П. (1999). Выбор данного словаря объясняется тем, что в дальнейшем в исследовании использовалась компьютерная программа eLex, предназначенная для составления словарных статей на основе значений слов именно по МАС.

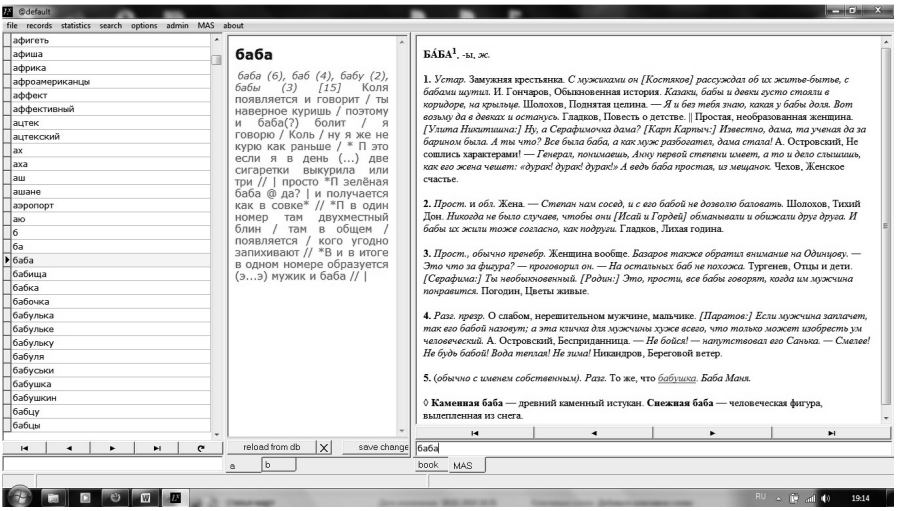


Рис. 1. Окно программы eLex, позволяющей сопоставлять данные ОРД и МАС

Ряду лексем из данных групп уже посвящены отдельные исследования. Так, например, было доказано, что в современной устной русской речи параллельно существуют две противоположные коннотации слова *девка*, ранее не зафиксированные в словарях: негативная (‘девица или легкого поведения, или малообразованная, из нижних слоев общества или неблагополучной семьи’) и положительная (‘хорошая, искренняя, прямая девушка’) (см. Осьмак 2011). Новую коннотацию в разговорной спонтанной речи получила и лексема *старушка*. Традиционно толковые словари определяют *старушку* как женщину, достигшую старости. Однако в ОРД встречается употребление данной лексемы в качестве обращения-обзывания:

- ***старушка*** *// сама ты* ***старушка*** *// поэтому …* ***старушка***(?) */ да*(?)

Анализ лексем *девочка/девчонка* и *мальчик* показал, что у них появились новые значения, которые можно охарактеризовать следующим образом:

1) *девочки/девчонки* — ‘лицо женского пола одного возраста с говорящим или младше’; характерно преимущественно для женской речи, а также для речи определенных возрастных групп мужчин (до 25 и после 60 лет);
2) *мальчик* — а) ‘лицо мужского пола одного возраста с говорящим или младше’, характерно только для женской речи; б) ‘лицо мужского пола младше говорящего’, характерно для мужской речи (Осьмак 2012).

Случаи особой сочетаемости лексем *дядя* и *тетя* (по преимуществу с именем собственным) привели к появлению таких новых фразеологизмов, как *тетя Мотя* (см. о подобных сочетаниях слов *дядя* и *тетя*: *Богданова* 2010б, в, 2011в).

Отдельный интерес вызывают лексемы, выявленные в ОРД, но не представленные в МАС. В основном это типичные диминутивы от *бабуля*, *бабуся, мужик* и *мужчина*: *бабулька*, *бабуська*, *мужичок* и *мужчинка*, ср.:

- *я говорю конечно / если она с ребёнком не умеет разговаривать / что это / **бабулька**-то ?*

- *там сидят **бабуськи** такие / на остановке // у них глаза такие по полтиннику // ну представляешь / они сидят / ждут автобуса;*

- *есть тут жилые дома вообще ? (э-э) дак (э) никто не знает // а потом () какой-то вот **мужичок** мне указал / что вот вам туда надо;*

- *ну как всегда у неё // достаёт классный телефон / тут ни с того ни с сего / начинает мне говорить / вот этот **мужчинка** / я уже не раз от неё такое слышала / она вот () рассказывала // да?*

Из примеров видно, что ничего принципиально нового в значениях этих слов, по сравнению с исходными, ОРД не дает. Различия могут быть, по-видимому, только стилистические или коннотативные.

Исключение представляет лексема *мужчинка*. Результаты блиц-опроса носителей языка (обращения к их ментальному лексикону) показали, что в настоящее время под *мужчинкой* понимается ‘недомужчина’, т. е. мужчина, по каким-то характеристикам недотягивающий до настоящего мужчины. Он может быть социально неуспешным, внешне несимпатичным или инфантильным, маленького роста, плюгавеньким. Однако нельзя сказать, что отношение носителей языка к *мужчинке* исключительно пренебрежительное. Часто данное слово употребляют в ироническом значении, даже с какой-то долей симпатии. В то же время все чаще встречается употребление лексемы *мужчинка* для обозначения представителей сексуальных меньшинств. Кроме этого, большинство носителей языка отмечают характерность употребления данного слова в женской речи. Богатейший материал для исследования значения этой лексемы предлагает и Интернет, в котором встречаются целые эссе-размышления о том, кто же такой *мужчинка*. Например:

- ***Мужчинка** — женское слово, редкий в русском языке пример однополого употребления. Это слово используют женщины на войне полов. Его феминистские корни не очевидны. Скорее, это взгляд сверху, лексикон высокомерия. Кто это там идет? Чего-то я не вижу… — **Мужчинка.** Плюгавенький такой, безденежный этакий. Что с него взять? Он ни на что не годится. С ним в ресторан пойти стыдно, но можно, хотя зачем? В кровать — ни за что. У него все маленькое, никудышное, жиденькое, нефотогеничное. Голый **мужчинка** — это, девки, обхохочешься!*

- *Бывают такие кадры, которых иначе и не назовёшь. Хиленькие, беспомощные, ведро воды на 2–3 метра не пронесут, не то чтоб девушку, маменькины сыночки — шагу в сторону без разрешения не сделают, того и гляди маму под окошком будут звать, такого на ручки разве что взять и пожалеть. И такие довольно часто встречаются, между прочим. Как их иначе назвать? — хлюпик? лучше* ***мужчинка****. И вроде обидно (должно быть для нормального мужчины) и без прямых оскорблений.*

- *Слово замечательное! Есть «мужчина» — это тот, за кем как за каменной стеной, добытчик и защитник. А есть «****мужчинка****» — такое нечто, мужеского пола, маменькин сынок до старости…*

Все эти примеры подтверждают необходимость более углубленного изучения данной лексемы.

Таким образом, становится очевидным, что работа с отдельными лексико-грамматическими группами слов дает богатейший материал для анализа семантических, коннотативных и даже стилистических изменений в функционировании слова в речи. В большинстве случаев для составления полноценной словарной статьи требуется дополнительное обращение к ментальному лексикону носителей языка через серию экспериментов в форме опроса — например, электронного — большого числа носителей русского языка, однако несомненно, что подобный лексикографический анализ следует провести в отношении и других лексико-грамматических групп, что и позволит, в конечном счете, составить ***словарь русской бытовой речи.***

Литература

1. *Богданова Н. В.* (2010а) Частотный словник ОРД — взгляд с «высоты птичьего полета». Материалы XXXIX Международной филологической конференции. Выпуск 23. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. СПб., 3–13.
2. *Богданова Н. В.* (2010б) Кто Вы, тетя Мотя? Мир русского слова, 4. СПб., 53–62.
3. *Богданова Н. В.* (2010в). Тети Моти, дяди Вади и проблемы перевода и преподавания русского языка. HUMANIORA: LINGUA RUSSICA. Труды по русской и славянской филологии. Лингвистика XIII. Развитие и вариативность языка в современном мире, I . Тарту, 9–26.
4. *Богданова Н. В.* (2011а) Об одном типе хезитационной вставки в русской устной спонтанной речи. Языковые категории и единицы: синтагматический аспект. Материалы девятой международной конференции. Владимир, 50–56.
5. *Богданова Н. В.* (2011б) Это самое: грамматические формы и функционирование в русской спонтанной речи. Материалы XL Международной

филологической конференции. Выпуск 24. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. СПб., 18–47.

6. *Богданова Н. В.* (2011в) Тетя Мотя в зеркале русской фразеологии. “WORT — TEXT — ZEIT”. Die slawische Phrazeologie in onomasiologischer, linguokultureller und phrazeographischer Sicht. Greifswalder Beiträge zur Slawistik X . Greifswald, 23–24.
7. *Богданова Н. В.* (2012) Про не это самое. Слово. Словарь. Словесность: Литературный язык вчера и сегодня (к 300-летию со дня рождения М. В. Ломоносова). Материалы ежегодной Всероссийской научной конференции. СПб., (в печати).
8. *Богданова Н. В., Бродт И. С., Куканова В. В., Павлова О. В., Сапунова Е. М., Филиппова Н. С.* (2008) О «корпусе» живой речи: принципы формирования и возможности описания. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог 2008». М., 57–61.
9. *Богданова Н. В., Осьмак Н. А.* (2011) О некоторых лексических «открытиях» на материале русской спонтанной речи (корпусное исследование). Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог 2011». М., 110–124.
10. *Бокале П.* (2011а) Дискурсивные слова в разговорной речи // russian.slavica.org/article9375.html.
11. *Бокале П.* (2011б) Роль текстовых коннекторов в молодежной речи // russian.slavica.org/article514.html.
12. *Герд А. С.* (1990) Новый академический словарь — исходные позиции и ориентиры. Актуальные проблемы разработки нового академического словаря русского языка. Л., 29–30.
13. *Дараган Ю. В.* (2011) Функции слов-«паразитов» в русской спонтанной речи. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Междугородней Конференции «Диалог 2000». // http://www.dialog-21.ru/materials/archive.asp?id=6260&vol=6077&y=2000
14. *Евгеньева А. П.* (1999) Словарь русского языка. В 4-х т. М.
15. *Кислощук А. И.* (2011) Опыт систематизации функций слова «вот» в спонтанном монологе. Актуальные проблемы современной науки. Материалы международной научной сессии «XIII Невские чтения». СПб., 157–164.
16. *Лаптева О. А.* (1999) Живая русская речь с телеэкрана. М., 520с.
17. *Осьмак Н. А.* (2011а) Из истории лексикографического описания русской спонтанной речи (далеко ли до Даля?). Материалы XL международной филологической конференции. Выпуск 24. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. СПб., 160–182.
18. *Осьмак Н. А.* (2011б) Новые коннотации старых слов (о лексикографическом описании русской спонтанной речи). Образовательные технологии в виртуальном лингвокоммуникативном пространстве. IV Международная виртуальная конференция по русистике, литературе и культуре. Вермонт, 166–170.

19. *Осьмак Н. А.* (2012) Новые значения старых слов (корпусное исследование на материале повседневной русской речи). (в печати).
20. *Степанова С. Б., Асиновский А. С., Богданова Н. В., Русакова М. В., Шерстинова Т. Ю.* (2008) Звуковой корпус русского языка повседневного общения «Один речевой день»: Концепция и состояние формирования. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог-2008». М., 488–494.
21. *Хан Н. А.* (2012) Конструкция (я) не знаю в русской спонтанной речи: соотношение разных функциональных типов. Мир русского слова. СПб., (в печати).
22. *Meyer C. F.* (2004) English Corpus Linguistics: An Introduction. Cambridge.

К СЕМАНТИКЕ КОММУНИКАТИВНОЙ СТРУКТУРЫ: ИСХОДНЫЕ СТРУКТУРЫ И ЛИНЕЙНО-АКЦЕНТНЫЕ ПРЕОБРАЗОВАНИЯ

Падучева Е. В. (elena.paducheva@yandex.ru)

ВИНИТИ, Российская академия наук, Москва, Россия

Планом выражения для коммуникативной структуры (К-структуры) является линейно-акцентная структура (ЛА-структура), т.е. линейная последовательность тоновых групп с определенными лексико-синтаксическими характеристиками. К-структура выявляет смысловые отношения между группами, возможно, иерархические. Ставится задача исчисления К-структур русского предложения и описания вклада К-структуры в семантику предложения. Задача решается в рамках трансформационного подхода. Выявлены исходные К-структуры (которые находятся в наиболее естественном соответствии с синтаксической структурой и лексическим составом предложения) и предъявлен набор ЛА-преобразований, которые порождают новые К-структуры. Для каждого преобразования указан его вклад в коммуникативную семантику предложения. Семантическая интерпретация К-структуры опирается на противопоставление presupposition – assertion.

Ключевые слова: коммуникативная структура, тема, рема, пресуппозиция, ассерция

1. План выражения и план содержания коммуникативной структуры

Коммуникативная структура предложения имеет план выражения и план содержания. Планом выражения для коммуникативной структуры (К-структуры) является **линейно-интонационная** структура предложения (Падучева 1985: 108); она же — **линейно-акцентная** структура (Янко 2001: 16). Линейно-акцентная структура (ЛА-структура) задается порядком слов и фразовой просодией, взятыми в их взаимодействии.

Планом содержания для К-структуры является членение предложения на коммуникативно релевантные (= **коммуникативно охарактеризованные**) компоненты и отношения между ними. До недавних пор единственными коммуникативно охарактеризованными компонентами предложения были **тема** и **рема**. А единственным семантическим ресурсом при описании семантического вклада коммуникативной структуры в семантику высказывания были понятия **логический субъект** и **логический предикат**. Иными словами, в классическом репертуаре было только одно отношение между компонентами

коммуникативного членения — субъектно-предикатное. Современная теория К-структуры располагает более развитым концептуальным аппаратом.

ЛА-структура — это последовательность тоновых (= акцентных) групп. К-структура опирается не только на ЛА-структуру предложения, но и на его лексико-синтаксические параметры.

2. Коммуникативная парадигма и ЛА-преобразование

Коммуникативная структура предложения связана его с лексико-синтаксической структурой (т. е. с синтаксисом и лексическим составом). Но не предопределена ею однозначно. На базе одной и той же лексико-синтаксической структуры можно построить целую **парадигму** коммуникативных структур (Падучева 1985: 108).

В развитие идей, заложенных в Ковтунова 1976, разработан подход к описанию коммуникативной семантики русского предложения, который включает набор **линейно-акцентных преобразований** коммуникативной структуры (Падучева 1985: 118, Циммерлинг 2008). Среди коммуникативных структур, образующих К-парадигму предложения, имеется **исходная**, или первичная, и **вторичные**, производные.

Вторичные К-структуры описываются с помощью линейно-акцентных преобразований исходных структур. ЛА-преобразование — это средство описания формально-смысловых соотношений между членами К-парадигмы. Это трансформационный подход к описанию семантического вклада К-структуры в семантику предложения. Он дает возможность описать варьирование порядка слов и интонации, указав изменения коммуникативной семантики предложения, которыми оно сопровождаются.

Основные исходные К-структуры являются **нейтральными** — в том смысле, что они находятся в наиболее естественном соответствии с синтаксической структурой и лексическим составом предложения: если К-структура нейтральная, то ее вклад в семантику предложения минимален или даже равен нулю. Однако есть исходные К-структуры, которые не являются нейтральными (а напротив, имеют богатую коммуникативную семантику), но не могут быть получены трансформацией никакой нейтральной структуры.

3. Коммуникативно охарактеризованные компоненты

Традиционные компоненты коммуникативной структуры — это тема и рема. Иногда различаются тема (= то, о чем говорится в предложении или тексте) и **кулиса**. И то, и другое может быть **началом** предложения (Падучева 1985: 113), так что в предложении может быть несколько начал. Так, в предложении (1) два начала — *рано утром* и *пионер Петя*:

(1) Рано утром пионер Петя открыл калитку.

В **нерасчлененном** (= тетическом) предложении нет ни темы, ни кулисы, т. е. нет начала. Термин тема употребляется в двух смыслах: в семантическом (ср. смысловая тема в разделе 5.4) и в чисто просодическом, когда он синонимичен термину начало. В этом смысле *однажды* может быть названо темой предложения (2):

(2) Однажды Пушкин прочёл Кипренскому стихи об Италии
[К. Г. Паустовский. Орест Кипренский (1936)]

Чтобы дать определение понятиям темы и ремы, необходимы некоторые сведения о просодии. В основе акцентов, которые играют смыслоразличительную роль в русских ЛА-структурах, лежат интонационные конструкции Е. А. Брызгуновой, см. Грамматика 1980: 99ff. Следуя за Янко 2001, мы принимаем следующие обозначения акцентов.

↘ — обычный понижающийся тон, обычно конечный, ИК-1,
◹ — усиленный (он же контрастный) понижающийся тон, ИК-2,
↗ — обычный повышающийся тон, ИК-3,
↱ — необычный повышающийся тон, ИК-6 (см. Грамматика 1980, т.1: 116).

Прочие акценты в дальнейшем изложении не используется. К этому списку следует добавить **структурно нерелевантное** понижение тона на конечном слове конечной составляющей, которое эквивалентно безударности; обозначается ___ . Например: *Совесть◹ надо иметь* ___ .

В любом предложении есть **главное фразовое ударение**, в повествовательном предложении нисходящее — ↘ или ◹, в вопросительном — восходящее. В расчлененном предложении есть также **второстепенное фразовое ударение**, обычно восходящее. Главное фразовое ударение обычно падает на конечное слово составляющей.

На каждый коммуникативно охарактеризованный компонент ЛА-структуры приходится один акцент. Слово, на которое падает акцент, — это **акцентоноситель**.

Теперь можно сказать, что **ремой** предложения является составляющая, в которую входит слово — носитель главного фразового акцента. А **тема** — это составляющая, которая отмечена (или может быть отмечена) второстепенным фразовым ударением. Вклад темы и ремы в семантику предложения описывается только в составе коммуникативной структуры, см. ниже.

Принципиальная связь нисходящего тона с главным фразовым ударением и восходящего — с второстепенным (так наз. интонация незавершенности) имеет место в независимом контексте, так сказать, в сильной позиции. При синтаксическом усложнении предложения происходят автоматические ЛА-преобразования (Падучева 1989; ср. также Подлесская 2008). Например, нисходящий рематический акцент заменяется на восходящий в контексте сочинения или условной конструкции:

(3) Пионер Петя открыл калитку↗ и вышел на большую зеленую лужайку↘ ;

(4) Если Петя открыл калитку↗, в сад забегут собаки↘ .

Коммуникативная структура может быть иерархической. Например, в (5) составляющая-тема в свою очередь членится на тему и рему (иерархия может возникать только в К-структуре, на базе синтаксиса; ЛА-структура — это всего лишь последовательность акцентных групп):

(5) [[Скорая помощь] [приехала]] [удивительно быстро]].

Ставится задача инвентаризации К-структур русских предложений. Описание состоит а) из набора исходных К-структур, где каждая охарактеризована соответствующей ей ЛА-структурой и **семантической интерпретацией**, см. раздел 4; и б) набора линейно-акцентных преобразований, которые меняют исходную ЛА-структуру и, соответственно, интерпретацию, вводя тем самым новые К-структуры, обеспеченные семантической интерпретацией, см. раздел 5.

4. К-структуры с исходным соотношением между синтаксическим и коммуникативным членением

Имеются следующие типы К-структур с исходным соотношением между синтаксическим и коммуникативным членением (Ковтунова 1976: 64).

4.1. Субъектно-предикатное расчленение

Это классическое расчленение на логический субъект и логический предикат. Тема (логический субъект) — именная группа, рема (логический предикат) — группа сказуемого:

(а) Октябрина Михайловна↗ | умерла↘ . <Вчера вечером.>
[А. Геласимов. Нежный возраст (2001)]

(б) Свою тещу↗ | он нежно любил↘ .

4.2. Детерминантно-событийное расчленение

Событийный компонент может быть нерасчлененным, как в (в), или членироваться на тему и рему, как в (г):

(в) 25 сентября 1970 года | умер Эрих Мария Ремарк.

(г) После тяжелой и продолжительной болезни | писатель умер.

4.3. Нерасчлененная структура (чистая рема; тетическое предложение)

(д) Умер Гавел↘ .

В нерасчлененном предложении нормой является субъектно-предикатная инверсия. В Падучева 2008 это объяснено тем, что препозиция подлежащего неминуемо влечет расчленение на тему и рему. В предложении (г), с детерминантно-событийным расчленением, такая же инверсия осуществляется в реме.

Односоставные предложения по определению нерасчлененные, и отдельных коммуникативных типов, вопреки Ковтунова 1976, не составляют: в предложении *На улице быстро темнело* расчленение типа 4.2, в *Становилось темно* — типа 4.3.

Во всех трех типах К-структур составляющие могут иметь более сложную синтаксическую структуру. Как мы видели, при одной реме может быть несколько тем, пример (1) из раздела 3. Аналогично, при одной теме может быть несколько рем (Падучева 1985: 113):

(е) Умер↗ Ремарк | 25 сентября 1970 года↗ | в возрасте 72-х лет↗ | в городе Локарно↘ .

Имеются две К-структуры, которые не могут быть получены ЛА-преобразованием никакой исходной структуры, хотя сами не являются нейтральными, а напротив, имеют богатую коммуникативную семантику. Это структура с неингерентной темой и с суперпозицией ремы.

4.4. Структура с неингерентной темой

Среди нерасчлененных предложений обратили на себя внимание предложения с главным фразовым ударением отнесенным в начало фразы (а не в конец, как обычно):

(1) <Тише.> Бабушка↘ спит.

(2) <Пустите!> Я на поезд↘ опаздываю.

В Баранов, Кобозева 1983 было введено понятие **неингерентной темы**. Природа предложений с неингерентной темой раскрыта в Янко 2001: 182–196: «ЛА-контур плюс выбор акцентоносителя <…> выполняют функцию опущенных предикатов пропозициональной установки». Фактически предложение с неингерентной темой представляет собой «разговорный вариант подчиненной пропозиции.» Например:

(3) Худо, брат, жить в Париже: есть↘ нечего (= ‘потому что есть нечего’).

Таким образом, в примере *Бабушка↘ спит* акцент на подлежащем нельзя считать результатом ЛА-преобразования «Экспрессивная препозиция ремы» (см. раздел 5.1) крайне неестественного предложения *Спит бабушка↘*. Инициальное ударение в (1) — это специальное оформление **предложения обоснования**.

Акцентоноситель в предложении с неингерентной темой выбирается по формальному правилу, так что «семантической нагрузки выбор того или иного слова не несет», Янко 2001: 182. Усматривается иерархия членов предложения с точки зрения их вероятности стать акцентоносителем. Ответ на вопрос *В чем проблема?* может быть *Едим↘ плохо*, *Дети↘ плохо едят* и *Дети кашу↘ плохо едят*.

Порядок слов, отличающий предложение с неингерентной темой от просто нерасчлененного, может порождать семантический эффект **топикализации** именной группы, выделенной главным фразовым акцентом — при том, что, согласно ЛА-структуре, темы в предложении нет:

(4) <Ну, слава богу, приехали. Опять ты дома. (Ласкаясь.)> Душечка↘ моя приехала! [А. П. Чехов. Вишневый сад (1904)]

Однако этот эффект не обязателен, ср. знаменитую фразу из той же драмы Чехова «Вишневый сад» (как ясно из контекста, нерасчлененную: *Епиходов* — новое), в которой ИГ-дополнение уводится из конечной позиции, но не с целью топикализации, а скорее чтобы избежать рематического акцента, чреватого контрастом:

(5) Яша (едва удерживаясь от смеха). Епиходов биллиардный кий↘ сломал!.. (Уходит.)
Варя. Зачем же Епиходов здесь? Кто ему позволил на биллиарде играть? Не понимаю этих людей... (Уходит.)

Таким образом, предложения с неингерентной темой — это специальная **акцентно оформленная** разновидность нерасчлененных предложений (в Николаева 1982: 65ff этот акцент был интерпретирован как выражение «глобальности» высказывания). Предложение *Идет снег↘*, нерасчлененное, может иметь и акцентно оформленную разновидность: *Снег↘ идет*; для нерасчлененного *Приехала моя душечка↘* это будет (4). А для предложения примера (5) позиция акцентоносителя — это единственное возможное выражение нерасчлененности. То же верно для (1).

4.5. Структура с суперпозицией ремы

Этот тип коммуникативных структур описан в Янко 2001: 161ff. Пример:

(6) Говорящего попугая↘ приобрел на Птичьем рынке студент Порошков↘ .

В предложении (6) по крайней мере две ремы, которые синтаксически не гомогенны, как в предложении (e) раздела 4.3, и потому не могут обе находиться в постпозиции к теме: одна — это агенс *студент Порошков*, другая — вводимый в рассмотрение объект, *говорящего попугая*.

В Циммерлинг 2008 К-структуру с суперпозицией ремы предлагается трактовать как результат применения преобразования Экспрессивная препозиция ремы (см. ниже раздел 5.1) к К-структуре, где уже есть коммуникативное членение на тему и рему. Т.е. (6) может быть получено из (6'):

(6') Студент Порошков↗ приобрел на Птичьем рынке↗ говорящего попугая↘ .

Предложение (7') (пример из Циммерлинг 2008) даст, в результате того же преобразования, предложение (7):

(7) Сильный взрыв↯ прогремел сегодня↗ на севере Британской столицы↘ .

(7') На севере Британской столицы↗ прогремел сегодня↗ сильный взрыв↘ .

Будучи сдвинуто со своего естественного места в конце фразы, главное фразовое ударение усиливается, так что акцент на сдвинутой группе не ИК-1, ↘ , а ИК-2, ↯.

5. Линейно-акцентные преобразования К-структур

Теперь обратимся к К-структурам с неисходным соотношением между синтаксическим и коммуникативным членением, в которых коммуникативный вклад описывается как результат ЛА-преобразования.

5.1. Экспрессивная препозиция ремы

Это ЛА-преобразование переносит собственно рему в начало предложения, сохраняя неизменным свойственный ей нисходящий фразовый акцент (Ковтунова 1976: 98). Понижение тона на последнем слове не имеет фонологического (т. е. смыслоразличительного) статуса.

(1) Вдруг странный треск↘ послышался в лесу ___
⇐ Вдруг в лесу послышался странный треск↘ .

В Янко 2001 показано, что это преобразование часто применяется в предложениях **введения в рассмотрение**.

В русском синтаксисе есть понятие **бытийное** предложение, см. Арутюнова 1976. Например, бытийные предложения играют роль в морфосинтаксисе отрицания. Но уже в связи с проблематикой отрицания выяснилось, что

бытийность дает слишком узкий класс — прежде всего потому, что имеется трудно прочерчиваемая граница между существованием и восприятием.

Описание коммуникативной структуры тоже выявило необходимость в классе, который связан с существованием, но является более широким, чем существование.

В предложениях введения в рассмотрение с ремой-подлежащим экспрессивная препозиция ремы дает эффект **топикализации ремы**.

Пример ниже трактуется в Янко 2001: 177 как предложение с суперпозицией ремы:

<… умирает Шевелев, полковой командир.> Женщина↘ сидит у его ног —— .

Однако это обычная экспрессивная препозиция ремы — которая в предложении введения в рассмотрение дает эффект топикализации. За пределами предложений введения в рассмотрение «эмфатическая» препозиция ремы не дает никакого специального коммуникативного отличия от предложений с конечной ремой:

Жаркие↘ были *споры* ⇐ *Споры* были жаркие↘ ;
Не однажды↘ вставал передо мною этот *вопрос* ⇐ Этот *вопрос* вставал передо мною не однажды↘ .

5.2. Дислокация ремы с рецессией темы

В Ковтунова 1976: 120 описаны два ЛА-преобразования, которые характеризуются как **инверсия с дислокацией ремы**: рема, состоящая из глагола и зависимой от него синтаксической группы, расчленяется на две части, и глагол выносится в начало предложения (в позицию перед подлежащим), а оставшаяся синтаксическая группа — в конец. При этом тема предложения заключается внутрь расчлененной ремы. Глагол получает (см. Янко 2001), акцент типа ИК-6, а оставшаяся синтаксическая группа сохраняет свой обычный главный фразовый акцент.

Коммуникативный эффект дислокации существенно зависит от того, было ли темой исходного предложения подлежащее или детерминант.

Предложение с темой-подлежащим при дислокации ремы дает, согласно Янко 2001: 201, структуру с **рецессией темы**. Тема подавлена, предложение уже не делится на тему и рему — оно становится нерасчлененным:

Посадил↦ *дед* репку↘ ⇐ *Дед*↗ посадил репку↘;
Пустился↦ *мой доктор* в разговоры↘ (пример из Падучева 1985).
Выйду↦ *я* за Ганса замуж↘ . Пойдет↦ *он* в погреб за пивом↘ .

Важное свойство этих предложений — неавтономность: типичная позиция такого предложения — в начале рассказа или анекдота; оно предполагает продолжение. Фраза *Плеханов ушел в кусты* (из Краткого курса истории

ВКП(б)) оказывается после этого преобразования неуместной в своем контексте, поскольку становится интродуктивной, т. е. требует продолжения:

Ушел↦ Плеханов в кусты↘ .

В К-структуре с дислокацией ремы тема находится, согласно Янко 2001, в «ваккернагелевской позиции» — после первого ударного слова; ее можно назвать **энклитической темой**.

5.3. Дислокация ремы в интродуктивном предложении

Если же в исходном предложении началом был детерминант, а подлежащее — собственно рема, то при той же дислокации ремы вокруг темы подлежащее оказывается в сильноударном положении, и результатом дислокации будет **интродуктивное** предложение (Арутюнова 1976: 221):

Служил↦ на Кавказе *один барин*↘ (пример из Янко 2001: 200),
Был↦ у майора Деева *товарищ*↘ (пример из Падучева 1985: 119).

Эти предложения неавтономны, как и предложения, в которых дислокация ремы сопровождается рецессией темы: и те и другие требуют продолжения. Но коммуникативный смысл их противоположен. Здесь подлежащее оказывается в рематической позиции, и соответствующий объект не затушевывается, а, прямо наоборот, вводится в рассмотрение.

5.4. Тема-рематическая инверсия; эффект идентификации

Одна из первых коммуникативных парадигм была предложена в статье Kuno 1972, где были рассмотрены четыре возможных акцентных варианта (и, соответственно, четыре К-структуры) для английского предложения *John kissed Mary* — кроме нейтральной интерпретации, рассматривался вариант с темой, с контрастом и с **интерпретацией исчерпывающего перечисления**. Представляет интерес эта последняя:

John↘ kissed Mary [exhaustive list interpretation] = ‘John (and only John) kissed Mary’.

В качестве возможной интерпретации для этого предложения рассматривалась также перефразировка ‘it was John who kissed Mary’. Перейдем к русскому примеру.

(1) а. Джон↗ | поцеловал Марию↘ [нейтральная К-структура];
б. Марию↗ поцеловал | Джон↘ [К-структура с тема-рематической инверсией].

Преобразование состоит в том, что тема и рема меняются местами: бывшая тема становится ремой (быть может, контрастной), а вся рематическая составляющая — темой. Дополнительная семантика — исчерпывающее перечисление: ‘только Джон поцеловал Марию’.

В Падучева 1985: 120 в описании семантики этого ЛА-преобразования было использовано понятие презумпции. Предложение (1а) отличается от (1б) тем, что в его семантику входит **презумпция** ‘существует <единственный> человек, который поцеловал Марию’ и идентифицирующая **ассерция**: ‘этот человек — Джон’.

Для того, чтобы данное ЛА-преобразование имело указанный семантический эффект, рема исходного предложения должна задавать идентифицирующее свойство объекта, обозначенного темой. Это условие выполняется в (2), и ЛА-преобразование сопровождается нужным семантическим эффектом:

(2) Вася меня проводит↘ ⇒ Проводит↗ меня Вася↘ = ‘существует <единственный> человек, который меня проводит, и этот человек — Вася’.

В примере (3) условие не выполняется, и результат соответствующего ЛА-преобразования с трудом поддается осмыслению (ср. Kuno 1972):

(3) а. Джон любит свободу; б. Свободу↗ любит Джон↘ .

Та же коммуникативная семантика (возникновение презумпции единственности) выражается конструкцией примера (4), которая является аналогом клефт-конструкции английского языка:

(4) Кто меня удивляет, это моя бабушка = Удивляет↗ меня моя бабушка↘ .

Предложение *Моя бабушка меня удивляет*, исходное для (4), никак не выражает уникальности бабушки.

Обращение к понятию презумпции позволяет дать интерпретацию известному примеру из Sgall, Hajičova 1977; предложение (5а) очевидным образом истинное. Между тем (5б) мы скорее склонны признать ложным:

(5) а. <u>В Австрии</u>↗ говорят на немецком языке↘ ;
б. На немецком языке↗ говорят <u>в Австрии</u>↘ .

Фраза (5а) содержит информацию о том, на каком языке говорят в Австрии, — на немецком; а фраза (5б) утверждает статус Австрии как единственной страны, где говорят на немецком языке: ее коммуникативная структура порождает семантику исчерпывающего перечисления.

В Янко 2001: 211ff был выделен тип предложений, охарактеризованный одновременно с трех точек зрения:

1) с линейно-акцентной точки зрения это с предложения с глаголом-темой;

2) с логико-семантической точки зрения это предложения со структурой «признак (атрибут) объекта — имя объекта — значение признака»;
3) с точки зрения роли в тексте: предложение содержит рецессивную тему.

Например:
<Служил на Кавказе один барин.> Звали↗ *его* Жилин↘ .
< — Можно я вас провожу? — Нет, спасибо. > Проводит↗ *меня* Вася↘ .

Глагол-тема указывает **признак** предмета-темы, а собственно рема — **значение признака**. Следует подчеркнуть, что здесь речь идет о **смысловой теме**, иначе — теме текста (Падучева 1985: 116) — просодической темой является глагол.

Предложенный анализ имеет принципиальное значение в плане расширения семантического арсенала теории коммуникативной структуры: наряду с субъектно-предикатным расчленением, двучленным, предлагается использовать параметрическое расчленение — трехчленное, в котором помимо двух обычных К-компонентов, тема и рема, участвует третий — энклитическая тема.

Основная часть предложений с глаголом-темой — это предложения, полученные ЛА-преобразованием тема-рематическая инверсия. Именно тематизация глагола порождает и презумпцию существования, и значение идентификации.

В структуре с детерминантно-событийным расчленением (см. раздел 4) тема-рематическая инверсия не дает никаких смысловых приращений — изменение касается только «порядка обхода ситуации» по Ю. С. Мартемьянову:

Но *Аня*↗ умерла | две недели назад↘ [Андрей Волос. Недвижимость (2000)]
⇐ Но две недели назад↗ | умерла *Аня*↘ .

При этом в тематической части возможна рецессия смысловой темы:
Но умерла↗ *Аня* | две недели назад↘ .

Так что рецессия смысловой темы не является конституирующим признаком предложений какого бы то ни было логико-семантического типа. Это отдельное ЛА-преобразование в предложениях, где тема допускает субъектно-предикатное расчленение:

Живут↗ *строители* по всей трассе канала↘ ⇐ Строители живут по всей трассе канала.
Выглядит↗ *она* превосходно ⇐ Она превосходно выглядит.

5.5. Тематизация собственно ремы

Эмфатическая препозиция (раздел 5.1) переносит собственно рему в начало, сохраняя неизменным ее фразовый акцент. Между тем тематизация собственно ремы меняет акцент фронтируемой составляющей

на противоположный. При этом предложение сохраняет свою двучленную структуру: бывшая тема и остаток ремы становятся единой рематической составляющей. Тематизация собственно ремы обычно сопровождается инверсией подлежащего и сказуемого (Падучева 2008).

(1) а. Мой *брат*↗ | *живет* в Казани↘ ; б. В Казани↗ | *живет* мой *брат*↘ ;

(2) а. Академик Вавилов↗ | *посетил* Лабораторию Завадовского↘ ;
б. Лабораторию Завадовского↗ | *посетил* академик Вавилов↘ .

Тематизация собственно ремы — не то же, что тема-рематическая инверсия. Формальное различие состоит в том, что при тема-рематической инверсии в начало выносится вся рема, а при тематизации собственно ремы — только субстантивная часть; глагол остается в реме. Различие семантическое состоит в том, что отношение идентификации, которое определяет семантику коммуникативного сдвига 5.1, при тематизация собственно ремы не возникает, см. (1), (2).

Вернемся к примеру (5) из раздела 5.4, на котором был продемонстрирован эффект идентификации:

(5) а. В Австрии↗ | *говорят* на немецком *языке*↘ ;
б. На немецком *языке*↗ | *говорят* в Австрии↘ .

Формально, т. е. на уровне ЛА-преобразования, в нем тематизации подвергается только собственно рема. Откуда тогда значение идентификации? Дело в том, что в этом предложении «сверхтесная» смысловая связь между глаголом и именной группой. Поэтому неважно, что в ЛА-структуре глагол относится к реме; по смыслу он входит в тему. Глагол *говорить* является идентифицирующим признаком для тематической ИГ *на немецком языке* в (5) в той же мере, в какой *поцеловать* — для ИГ Марию в (1).

Субъектно-предикатную инверсию в примерах (1), (2) естественно объяснить тем, что обратный порядок слов служит средством выражения коммуникативной нерасчлененности — не только в целом предложении, но и в его субтематической части.

5.6. Контрастная тема

ЛА-преобразование состоит в том, что начальная группа с обычным (в том числе — факультативным) второстепенным ударением приобретает констрастное:

(1) а. Иван ее любит↘ ; б. Иван↗ ее любит↘ .

Содержательный сдвиг (Падучева 1985: 118) состоит в том, что вводится в рассмотрение множество объектов, **ассоциированное** с объектом-темой

a (в которое входит и сам объект *a*). А в смысл предложения, помимо утверждения о том, что некое обстоятельство Q (рема) имеет место для *a*, входит, в статусе импликатуры, компонент ‘для других элементов ассоциированного множества, возможно, Q не имеет места’. См. о множестве альтернатив в Rooth 1992 и о контрасте в Богуславский 1998. Так, в смысл (б) входит не только компонент (а), ‘Иван ее любит’, но и импликатура ‘и, возможно, больше никто’.

Есть серия кванторных наречий и прилагательных, которые недопустимы в позиции контрастной темы (Падучева 2005; см. о неконтрастоспособных словах в Янко 2001: 19ff).

(2) а. *Некоторые*↗ инструменты нам понадобятся↘ [= ‘а другие нет’];
б. **Все*↗ инструменты нам понадобятся↘ ;

(3) а. *Какой*↗*-нибудь* молоток найдется↘ [возможно, не лучший];
б. **Всякий*↗ молоток найдется↘ .

Такая избирательность легко выводится из определения контрастной темы. Слова *все*, *всегда*, *всякий* исключены в позиции контрастной темы, поскольку не порождают ассоциированного множества: *некоторые*↗ *люди* противопоставлены некоторым другим; а *все*↗ *люди* уже ничему не противопоставлены.

———

Можно думать, предлагаемый способ представления коммуникативной структуры является достаточной основой для описания вклада коммуникативной структуры в семантику русского предложения. В описании семантики К-структуры главное место занимает противопоставление презумпция — ассерция. Это позволяет К-структуре занять центральное место в семантике отрицательного предложения, внести ясность в семантику и синтаксис биноминативного предложения, прояснить семантику акцентных сдвигов в предложениях с адвербиалами, количественными конструкциями и многое другое. Но об этом — в следующий раз.*

Литература

1. *Апресян* 1974 — Апресян Ю. Д. Лексическая семантика: Синонимические средства языка. М.: Наука, 1974.
2. *Арутюнова* 1976 — Арутюнова Н. Д. Предложение и его смысл. М.: Наука, 1976.

* Автор искренне благодарен двум анонимным рецензентам. Их замечания и соображения, по силе возможностей, учтены в окончательном варианте текста.

3. *Богуславский* 1998 — Богуславский И. М. Сфера действия начинательности и актуальное членение: втягивание ремы // Семиотика и информатика. Вып. 36. М.: Языки рус. культуры: Рус. словари, 1998.
4. *Грамматика* 1980 — Русская грамматика. Т. 1. Отв. ред. Н. Ю. Шведова. М.: Наука, 1980.
5. *Ковтунова* 1976 — Ковтунова И. И. Современный русский язык: Порядок слов и актуальное членение предложения. М., 1976.
6. *Николаева* 1982 — Николаева Т. М. Семантика акцентного выделения. М.: Наука, 1982.
7. *Падучева* 1985 — Падучева Е. В. Высказывание и его соотнесенность с действительностью. М.: Наука, 1985.
8. *Падучева* 1989 — К интонационной транскрипции для предложений произвольной синтаксической сложности. Вопросы кибернетики. Семиотические исследования М.: АН СССР. Научный совет по комплексной проблеме «Кибернетика», 1989, 18–29/
9. *Падучева* 2005 — Падучева Е. В. Семантика количества и ее отражение в просодии. // Логический анализ языка. Квантификативный аспект языка. М.: Индрик, 2005, 652–667.
10. *Падучева* 2008 — Падучева Е. В. Коммуникативная расчлененность и пути ее преодоления: инверсия подлежащего. //Фонетика и нефонетика. М.: ЯСК, 2008, 417–426.
11. *Подлесская* 2008 — Подлесская В. И. Фразовая акцентуация в относительных предложениях: анализ корпусных данных //Фонетика и нефонетика. М.: ЯСК, 2008, 427–445.
12. *Циммерлинг* 2008 — Циммерлинг А. В. Локальные и глобальные правила в синтаксисе. Труды международной конференции Диалог 2008, 551–562.
13. *Янко* 2001 — Янко Т. Е. Коммуникативные стратегии русской речи. М.: Языки славянской культуры, 2001.
14. *Янко* 2008 — Янко Т. Е. Интонационные стратегии русской речи в сопоставительном аспекте. М.: Языки славянских культур, 2008.
15. *Kuno S.* Functional sentence perspective. A case study from Japanese and English. Linguistic inquiry v.3 N3, 1972, 260–320
16. *Rooth* 1992 — Rooth M. A theory of focus interpretation. Natural Language Semantics, 1992, v. 1, 1–116.
17. *Sgall, Hajičova* 1977 — Sgall P., Hajičova E. Focus on focus. I. — Prague bull. math. linguistics, v. 28, 5–54, 1977.

ПРОБЛЕМЫ И МЕТОДЫ АНАЛИЗА РУССКИХ ТЕКСТОВ В ДОРЕФОРМЕННОЙ ОРФОГРАФИИ[1]

Поляков А. Е. (pollex@mail.ru)

ГНПБ им. К. Д. Ушинского, Москва, Россия

Существующие лингвистические процессоры не пригодны для анализа текстов в дореформенной орфографии из-за многочисленных графических, морфологических и лексических отличий языка 18–19 века от современного языка. Мы разработали лемматизатор, который умеет правильно анализировать тексты в дореформенной орфографии, а также включает возможность гибкой настройки на другие орфографические системы (включая смешанную орфографию). В данной работе рассматриваются проблемы, возникающие при анализе русских дореформенных текстов, и возможные пути их решения.

Ключевые слова: словоизменение, русский язык, дореформенная орфография, автоматический анализ

1. Постановка проблемы

Исследователи, занимающиеся анализом текстов в старой орфографии, особенно текстов 18–19 века, хорошо знают, что существующие лингвистические средства (спеллчекеры, лемматизаторы, программы распознавания), ориентированные на современный язык, не годятся для анализа старых текстов. Эти программы правильно анализируют формы, совпадающие с современными (*рука, новый, милость*), однако формы, отличающиеся от современных (*домъ, домь, новаго, милостію, безсильныя, ходити*), не распознаются (считаются ошибкой), даже если все отличие сводится к конечному -ъ.

Лемматизатор mystem [1] имеет ограниченную поддержку старой орфографии путем приведения ее к современному написанию по простым формальным правилам (см. п. 3.1). Он правильно анализирует многие дореформенные написания, однако не может учесть все устаревшие формы и леммы и нередко строит для них неправдоподобные гипотезы. Для адаптации программы к языку 18–19 века необходима серьезная правка словаря и грамматических таблиц, однако код программы и словарей закрыт.

Лемматизатор Диалинг [2] и основанный на нем pymorphy [3] имеют открытый код, который можно адаптировать для старой орфографии. Однако

[1] Данное исследование выполнено в рамках работ по гранту РФФИ 11-06-00197 «Создание программного модуля проверки русской дореформенной орфографии».

сами словари и грамматические таблицы представлены в техническом формате, который непрозрачен и неудобен для редактирования, а преобразование их в читаемый вид потребует серьезной переделки всего кода.

Список других морфологических анализаторов для русского языка можно найти в материалах форума «Оценка методов автоматического анализа текста» [4]. Некоторые из этих систем, помимо лемматизации, способны решать более сложные задачи: порождение гипотез для нераспознанных слов, разрешение синтаксической неоднозначности (дизамбигуация) и др. Однако все они ориентированы на современный русский язык, а возможность их настройки на другие орфографические системы и модели словоизменения представляется достаточно сомнительной.

Вместо того, чтобы пытаться переделывать чужой код, нам оказалось проще написать собственный лемматизатор, включающий возможность гибкой настройки и адаптации к языку 18–19 века, который имеет заметные орфографические, морфологические и лексические отличия от современного языка. Ключевым компонентом для решения этой задачи являются не программы и алгоритмы, а полнота и точность словаря и грамматического описания, а также удобный формат для их записи.

В принципе, для анализа текстов вместо лемматизатора можно сделать простой список словоформ с разборами. Однако, чтобы сделать такой список, нужно вначале научиться разбирать слова, причем не вручную, а автоматически. Поэтому даже на начальном этапе все равно нужен какой-то лемматизатор, в который нужно ввести хотя бы минимальную информацию о словоизменении (иначе он просто не сможет работать), а дальше обучать. Такой метод является единственно возможным для малоизученных языков, для которых нет нормального словаря и грамматического описания, поэтому они создаются индуктивно в процессе анализа существующих текстов. Однако это не относится к русским текстам 18–19 века, для которых существует обширная словарная и грамматическая база.

2. Описание программы

2.1. Определения

Морфологический анализатор (лемматизатор) — программа, выполняющая грамматический разбор текста, который включает в себя следующие задачи:

1) токенизация — разбиение текста на элементарные знаки (токены) и определение типа для каждого токена: слово, знак препинания, цифровой комплекс, тег разметки и т. д.
2) морфологический анализ для слов, присутствующих в грамматическом словаре;
3) построение гипотез для нераспознанных слов (если это возможно).

Грамматическая модель — формальное описание словоизменения данного языка, которое включает два компонента: грамматический словарь и таблица парадигм.

Грамматический словарь — список лексем языка с приписанной информацией о словоизменении. Каждая лексема в словаре содержит, как минимум, следующую информацию:

1) основа с указанием чередований;
2) постоянные признаки лексемы (часть речи, род, одушевленность, переходность, и т. д.);
3) код словоизменительного типа (парадигмы).

Словоизменительный тип (парадигма) — набор флексий (с учетом схем чередования), общий для некоторого множества лексем. Словоизменительный тип задает соответствие между грамматическими значениями и соответствующими флексиями и чередованиями.

2.2. Принципы

Лемматизатор не привязан жестко к конкретному языку и в принципе может быть настроен на любой язык флективного типа при наличии соответствующей грамматической модели и правил токенизации. Вся конкретно-языковая информация (леммы, основы, парадигмы, флексии, граммемы) не зашита в код программы, а вынесена во внешние таблицы (словарь + парадигмы). Для записи словоизменительной информации была разработана специальная нотация, которую легко читать и править вручную.

Лемматизатор включает механизм, позволяющий ему адаптироваться к разным орфографическим системам. В процессе анализа исходный текст преобразуется во внутреннее (нормализованное) представление, которое является абстракцией от реального написания и может соответствовать нескольким орфографическим вариантам. Например, для анализа старых написаний приставок с ъ+гласная (*съиграть, съузить, съэкономить*) достаточно ввести правило замены *ъи=>ы, ъу=>у, ъэ=>э*. Механизм графических преобразований позволяет анализировать тексты с плавающей орфографией, где различные варианты написания синонимичны и свободно чередуются (*и/i/v*, *ф/ѳ*, *о/w*, *ъ/'*/пусто), хотя в целом анализ плавающей орфографии не сводится к простым правилам замены букв.

2.3. Алгоритм

Алгоритм работы лемматизатора состоит из следующих этапов:

1) Токенизация — программа разбивает текст на элементарные знаки (токены) и выделяет слова.
2) Нормализация — программа переводит слово во внутреннее представление, а исходная форма слова сохраняется для выдачи.

3) Программа пытается в цикле разбить слово на основу и флексию, а затем для каждого варианта разбиения проверяет дополнительные условия:
   - основа сочетается с данной флексией (имеет соответствующую парадигму);
   - основа имеет правильный вариант (при наличии чередования);
   - основа совместима с лексемой по регистру;
   - граммемы флексии совместимы с граммемами лексемы (см. п. 2.4).

   Для ускорения анализа используются заранее построенные таблицы флексий, основ и парадигм.
4) При отсутствии вариантов программа пытается построить гипотетический разбор по аналогии с существующими словами. При этом используется статистическая таблица характерных концов слова и их возможных разборов, а гипотезы сортируются в порядке убывания вероятности.

2.4. Грамматические фильтры

Правила сочетаемости граммем обычно зависят не от парадигмы, а от грамматического класса слова, и часто имеют очевидную семантическую мотивацию. Вот некоторые правила:

1) Глаголы сов. вида не имеют форм презенса (**сделающий,*сделаемый*), а формы, соответствующие презенсу несов. вида, имеют значение будущего (*делаю*=ind,pres vs. *сделаю*=ind,fut).
2) Непереходные глаголы не имеют форм пассива (*варимый, варенный* vs **веримый, *веренный*), кроме некоторых специфических конструкций (*хожено, сижено*), и не имеют медиальных форм (*-ся*), за исключением особой безличной конструкции (*не сидится, не лежится, не гуляется ему*).
3) Многократные глаголы (*сиживать*) не имеют форм презенса (**сиживаю*), а только претерита.
4) Безличные глаголы (*тошнить*) имеют только формы pres,sg,3 (*тошнит*) и past,sg,n (*тошнило*).
5) Относительные прилагательные (*вчерашний*) не имеют кратких форм и сравнительной степени.
6) Аккузатив совпадает с генитивом для одушевленных имен и с номинативом для неодушевленных, кроме некоторых парадигм (*сестру, лошадь, новую, его, их*).

Формальное наложение этих правил на парадигмы приводит к комбинаторному взрыву числа типов. Так, вместо одного типа V1 (*делать*) у нас появится 4–8–16 типов в зависимости от вида, переходности, многократности, безличности и других лексических признаков. Аналогично вместо одного типа N1 (*стол*) появится 2–4 типа в зависимости от одушевленности, лексического числа (sg/pl tantum) и др.

Механическое размножение парадигм крайне неестественно с лингвистической точки зрения и неэффективно с программной. Вместо этого в лемматизатор включены правила сочетаемости граммем, которые отсекают несовместимые комбинации или корректируют значения граммем (pres => fut).

2.5. Возможности и области применения

Созданный нами лемматизатор имеет довольно широкие возможности:

1) он умеет анализировать текст в разных орфографиях (современной, дореформенной, смешанной);
2) он умеет строить гипотезы для слов, отсутствующих в словаре;
3) он понимает различные входные форматы (включая текст с xml-подобной разметкой) и порождает несколько выходных форматов (разборы в скобках, xml-формат, табличный формат).

Лемматизатор может работать в нескольких режимах:

1) полная лемматизация, где выдаются все варианты разбора с грамматической информацией;
2) частичная лемматизация, где выдаются леммы, а грамматическая информация дается в свернутом виде или не дается вообще;
3) режим спеллчекера, где просто помечаются нераспознанные слова, но не дается грамматической информации и не порождаются гипотезы.

В настоящее время лемматизатор используется в следующих проектах:

- морфологический анализ для текстов 18 века для Национального корпуса русского языка;
- создание словарей и конкордансов к произведениям русских писателей (Ломоносов, Батюшков);
- подготовка текстов в старой орфографии для электронных библиотек;

и некоторых других.

3. Лингвистическая модель

3.1. Введение

При создании лемматизатора мы отталкивались от грамматической модели современного русского языка, зафиксированной в словаре Зализняка [5]. Мы изучили грамматические и орфографические описания 18–19 века (Ломоносов [6], Греч [7], Востоков, Грот [8]) для выявления основных особенностей языка данного периода.

Эти особенности можно условно разделить на три группы:

1) орфография (дополнительные буквы и другие правила написания);

2) словоизменение (устаревшие формы);
3) лексика (устаревшие слова).

3.2. Орфография (упрощенный вариант)

Поскольку современная орфография в основном является упрощением старой, сразу возникает идея, что старую орфографию можно анализировать путем приведения к современной. Вот основные правила преобразования старой орфографии в новую:

1) заменить устаревшие буквы на современные эквиваленты (*і=>и, ѣ=>е, ѳ=>ф, ѵ=>и*);
2) отсечь конечный *-ъ*;
3) заменить начальные *без-/в(о)з-/из-/низ-/раз-/роз-/ч(е)рез- => -с* перед глухими (NB);
4) заменить конечные *-аго/яго(+ся) => -ого/его*, *-ыя/ия(ся) => ые/ие* (NB).

Мы проверили этот метод на большом корпусе текстов в старой орфографии и получили неплохие результаты. Большая часть слов была разобрана правильно, неопознанными остались только устаревшие формы и лексемы.

Однако метод упрощения, хотя позволяет быстро получить результат, создает слишком много шума. Во-первых, механическая замена без учета морфологической структуры слова иногда работает неверно, например, слова *низкий, благо, стихия* заменяются на несуществующие формы **ниский, *блого, *стихие*. Во-вторых, смешиваются квази-омонимы, которые различаются именно буквами *ѣ–е* (*слѣзъ–слезъ, сѣлъ–селъ, свѣдѣніе–сведеніе, Вѣна–вена, морѣ–море*). В-третьих, невозможно отличить ошибочные написания, которые при переводе в современный вид совпадают с правильными (*ѣлка, ѳѵзика, істория*). Очевидно, что для правильного анализа старой орфографии необходим словарь и грамматическое описание, где точно указаны места употребления букв (прежде всего, *ѣ* и *ѳ*) в составе корней, суффиксов и флексий, а также учтены устаревшие формы.

3.3. Словоизменение

Если взять за основу модель словоизменения современного русского языка, то для анализа текстов в старой орфографии, помимо очевидных графических отличий, необходимо добавить в грамматическую модель формы, отсутствующие в современном языке или не учтенные в словаре Зализняка:

1) Адъективные флексии (*-аго/-яго, -ыя/-ія*).
2) Усеченные формы прилагательных (*красна/о/ы/у*), которые формально совпадают с краткими формами в именительном падеже, однако имеют и другие падежные формы (*красну*).
3) Особые формы местоимений (*ея, онѣ, однѣ, однѣхъ*).

4) Творительный падеж 3-го склонения на *-ію* (*милостію, помощію*).
5) Сравнительная степень на *-ѣй* (*сильнѣй*) и *-яе* (*сильняе, скоряе*).
6) Вариант частицы *-ся* после гласных (*валюся, валилася*), который употребляется в современном языке (в некоторых идиолектах).
7) Деепричастия совершенного вида от основы презенса (*прийдя, увидя, взгромоздясь*), которые вполне употребительны в современном языке, но не учтены в словаре Зализняка.
8) Глагольные флексии *-ти* и *-ши* (*ходиши, ходити*), которые употребляются в основном в 18-ом веке и, видимо, должны трактоваться как церковнославянизмы.

В грамматических таблицах некоторые из этих форм имеют специальные пометы (устаревшая, церковнославянизм), чтобы их можно было отличить от общих и современных форм.

4. Лингвистические особенности текстов

4.1. Введение

Характерной особенностью языка 18–19 века является значительная лексическая и орфографическая вариативность по сравнению с современным языком. Там, где в современном языке зафиксирован один лексический вариант, в старых текстах нередко встречается несколько взаимозаменяемых вариантов. Орфографические нормы также неоднократно менялись в течение 18–20 веков (ср. орфографию Ломоносова, Греча и Грота). Авторы часто не придерживались четкой системы, издатели исправляли авторскую орфографию, а в советское время многие тексты были переизданы в модернизированной орфографии. В результате в существующих текстах наблюдается причудливое смешение разных орфографических систем, с которым вынужден работать наш лемматизатор.

Подробное описание орфографических особенностей и колебаний в истории русского языка 17–19 века содержится в работах [9], [10], [11].

Ниже мы рассмотрим некоторые языковые и орфографические особенности текстов указанного периода и возможные методы и алгоритмы их анализа. Большинство примеров взято из поэтических текстов Ломоносова, опубликованных в 8-м томе Полного собрания сочинений, где сохраняются многие особенности авторской орфографии, хотя в частично модернизированной форме.

4.2. Лексические особенности и вариативность

1) Притяжательные прилагательные.

В современном языке притяжательные прилагательные образуются в основном от собственных имен первого склонения (*Петин, Машин + мамин,*

дядин). В языке 18–19 века они образуются регулярно от всех собственных имен и многих нарицательных. Примеры: *Августов, Алцидов, Ахиллесов, Бакхусов, Варронов, Венерин, Виргилиев, Гекторов, Енеев, Дафнисов, Екатеринин, Зевесов, Ликургов, Марсов, Минервин + государев, государынин, отцов, царев*.

Сейчас лемматизатор разбирает некоторые из этих слов то как фамилии, то как косвенные формы имен. Для адекватного анализа нужно включить эти формы в расширенную парадигму имени или ввести в словарь некоторое количество притяжательных форм для обучения модуля предсказания.

2) Вариативность приставок *воз/вз*.

В современном русском языке приставки *воз/вз* четко различаются — *восстать* vs. *встать*, *восход* vs. *всход(ы)*, *воспитание* (**вспитание*), *взлететь* (**возлететь*). Лишь в некоторых однокоренных словах есть варьирование: *вспоминать–воспоминание, возлюбить–невзлюбить, вздыхать–воздыхатель*.

В языке Ломоносова эти приставки, видимо, абсолютно синонимичны и чередуются свободно. Есть десятки примеров, где между вариантами невозможно найти какого-либо смыслового или стилистического различия: *возбудить–взбудить, возвеселить–взвеселить, возвести–взвести, возгордиться–взгордиться, воздеть–вздеть, воздыхание–вздыхание, воздыхать–вздыхать, возлететь–взлететь, возложить–взложить, возлюбить–взлюбить, возмужать–взмужать, вознести–взнести, возрастить–взрастить, воспевать–вспевать, воспомянуть–вспомянуть, воспылать–вспылать, воспятить–вспятить, восток–всток* (NB), *восточный–всточный, востребетать–встребетать, восход–всход, восходить–всходить*. Разумеется, это варьирование не абсолютно, некоторые слова имеют только один вариант, но все же довольно характерно.

3) Варианты суффикса *-ный/-ний*.

Примеры: *внутренний (-ый), внутренне (-о), всегдашний (-ый), дальный (-ий), дольный (-ий), искренний (-ый), крайний (-ый), осенний (-ый)*. В современном языке такое колебание сохранилось в слове *междугородний (-ый)* и в паре *искренний–искренно*.

4а) Варьирование *и/ъ* в конце основы.

Примеры: *видение (-ье), владение (-ье), возмущение (-ье), внимание (-ье), волнение (-ье), дыхание (-ье), желание (-ье), житие (-ье), здравие (-ье), копье (-ие), оружие (-ье), щастие (щастье)*.

В современных словарях обычно зафиксирован один вариант: *варенье* vs. *варение*, *беганье* vs. *плавание*. Однако в реальности такие варианты возможны почти для любого слова и воспринимаются как свободные колебания типа *водой–водою*. Эти варианты активно используются в поэтической речи для подстановки слова в различные ритмические контексты.

Для адекватного анализа таких вариантов мы добавили в словарь и в программу условный символ (гиперграфему), который может соответствовать обоим вариантам — *и/ъ*.

4б) Варьирование *и/ь*+гласная в середине основы.

Примеры: *азиатец–азиятец, италианский–италиянский–итальянский, баталион–батальон, материал–матерьял, вариант–варьянт*.

Это варьирование аналогично предыдущему, но лексически более ограничено, хотя некоторые колебания сохраняются и в современном языке. Оно похоже на характерное соотношение между церковнославянскими и русскими формами: *бию–бью, Татиана–Татьяна, Иулиан–Юлиан–Ульян*.

4.3. Орфографические особенности и вариативность

Основные точки вариативности в старой орфографии — те же, что и в современной:

1) обозначение мягкости шипящих (*меч* vs. *ночь*, *куш* vs. *идешь*) и др. согласных (*росте* vs. *бросьте*);
2) обозначение *ё/о* после шипящих и *ц* (*чёрт* vs. *чохом*, *шёлк* vs. *мешок*);
3) глухие/звонкие согласные в позиции нейтрализации (*воздать* vs. *воскресить*);
4) употребление прописные букв (*Англия* vs. *английский, англичанин*);
5) слитное/раздельное/дефисное написание (*тоже* vs. *то же*, *надвое* vs. *по двое*, *вмиг* vs. *в минуту*).

В истории русской орфографии правила написания в этих точках неоднократно менялись, а для случая 5 (слитно/раздельно) внятные правила не выработаны до сих пор.

Вот некоторые примеры орфографических колебаний в текстах Ломоносова и способы их решения:

1а) Написание ь/ноль после мягких согласных в основе:

вер(ь)вь, вер(ь)х, вет(ь)вь, пер(ь)вый, цер(ь)ковь, гор(ь)кий, сил(ь)ный, т(ь)ма, воз(ь)му.

Это в основном лексические колебания, касающиеся конкретных слов, но вообще факультативность написания ь внутри слова характерна для более раннего периода русского письма.

1б) Написание ь/ноль после шипящих во флексиях:

будеш(ь), видиш(ь), возбудиш(ь), восстаеш(ь), даеш(ь), держиш(ь), дерзнеш(ь), желаеш(ь), знаеш(ь), зриш(ь), идеш(ь), имееш(ь), любиш(ь), боиш(ь)ся, возмутиш(ь)ся.

бичь, ключь, лучь, мечь, Васильевичь, Ивановичь, Михайловичь;

Для анализа таких форм мы добавили флексии *-ш* и *-шся* в парадигму глагола, а также флексию *-ь* для имен на ч.

2) Написание *е/о* после шипящих и *ц*:

агньцов, венцев–венцов, венцем, зайцов, концев–концов, лисицой–лисицей, лицо–лице, лицом–лицем, любимцов, месяцов–месяцев, младенцов,

отцев–отцов, отцем–отцом, праотцов, пришельцов, пшеницой–царицей, свинцом–певцем, стрельцов–пловцев;

несчотный–несчетный, предпочол–предпочел, чорной–черной, чорны–черны, мужичок, сверьчок, старичок, плечо, счота, учоный, чорт — бичем;

большой–большей, большом–большем, шол–шел, вшол–вошел, обшол–обшел, отшол–отшел; поджог, рожок — душей, душею; чужом–чужем, чужою–чужею.

Колебание во флексиях можно решить алгоритмически, если ввести оба флексий (*о/е, ом/ем, ов/ев*) во все соответствующие парадигмы, независимо от места ударения. Колебание в основе можно решить, только если ввести «гиперграфему» *о/е* в соответствующие лексемы, однако в любом случае это будет словарное правило.

3) Прилагательные на *-зкий* (аналогия с частотными на *-ский — росский*): *близкий–блиский, дерзкий–дерзский–дерсский, мерзкий–мерзский–мерсский.*

Это индивидуальная особенность данных слов, которая не решается алгоритмически.

4) Прописная буква используется не только в собственных именах, но также в следующих случаях:

титулы — *Император, Самодержец, Величество + Генерал Порутчик, Академик, Кавалер* (иногда);

названия народов (не всегда) — *Немцы, Немецкий, Шведы, Шведский*;

производные от топонимов — *Азийский, Американский, Афинский, Багдатский, Берлинский*;

названия месяцев, наук, искусств, олицетворения и др.

Прописная буква не составляет проблемы для лемматизатора, поскольку в русском языке любое слово, нормально пишущееся со строчной, получает первую прописную в определенных контекстах. Однако это небезразлично для словаря, где нужно выбрать словарную форму слова, исходя из наиболее частотного написания, а также для спеллчекера, где регистр важен для проверки правильности слова.

5) Слитно или раздельно:

не+глаголы, наречия и др. — *небойся, недай, недайте, неокончав, неопасайся, неиначе, неможно*;

предлоги (редко) — *длятого, изовсего, вних*;

частицы *б(ы), ж(е), ли(ль)* — *былаб(ы), вернаяб, весьмабы, взойтиб, возможноли(-ль), вамиб, гдеб, гдеж, горыль, давноб, давноль, довольноль, долголь, Елисаветаб, естьлиб(ы), молвиж, моглиб, могуль, намиб, мыль, нашаб, нуж, ониб.*

Для анализа таких форм мы включили в программу дополнительное правило для слитных клитик (*не, бы, же, ли*), которое срабатывает, если прямой анализ не дал результата.

6) Написания *сч/щ, жч/щ*:

счастье–щастье, счастливый–щастливый, разсчет–расчет–ращет, мужчина–мущина.

Это индивидуальная особенность данных слов, которая не решается алгоритмически.

В целом, для анализа орфографической вариативности нет универсального метода, но можно предложить частные решения для конкретных случаев:

1) Нормализация — преобразование исходного написания в нормализованное, которое далее анализируется по стандартной грамматической модели.
2) Расширение грамматической модели — включение дополнительных форм в парадигмы.
3) Расширение лексической модели — специальная нотация для записи лексической вариативности («гиперграфемы») или включение дополнительных вариантов в словарь.
4) Модификация алгоритма для анализа слитных написаний.

5. Заключение

Для анализа текстов 18–19 века необходимо создание полноценной словарной базы, прежде всего, **грамматического словаря**, который содержит все основные лексемы и грамматические формы, характерные для языка данного периода. В качестве лексической базы для такого словаря нужно использовать статистические данные, полученные из реальных текстов, а также следующие лексикографические источники:

- Словарь Академии Российской (1789–1794);
- Словарь церковнославянского и русского языка (1847);
- Полный русский орфографический словарь (1898);
- Словарь русского языка 18 века;

и некоторые другие.

Литература

1. *Морфологический* анализатор mystem (http://company.yandex.ru/technologies/mystem/).
2. *Сокирко А. В.* Русский морфологический словарь (http://aot.ru/docs/rusmorph.html).
3. *Морфологический* анализатор Pymorphy (http://bitbucket.org/kmike/pymorphy/).
4. *Оценка* методов автоматического анализа текста: морфологические парсеры русского языка (http://ru-eval.ru/participants.html).
5. *Зализняк А. А.* Грамматический словарь русского языка. — М., 2008.

6. *Ломоносов М. В.* Российская грамматика. — СПб., 1788.
7. *Греч Н. И.* Практическая грамматика русского языка. — СПб., 1827.
8. *Грот Я. К.* Спорные вопросы русского правописания от Петра Великого доныне. — СПб., 1873.
9. *Ильинская И. С.* Лексика стихотворной речи Пушкина. «Высокие» и поэтические славянизмы. — М., 1970.
10. *Перцов Н. В.* О соотношении письменной и устной форм поэтического языка (К вопросу о функциональной нагруженности старого русского правописания) // Вопросы языкознания. 2008. № 2. С. 30–56.
11. *Каверина В. В.* Становление русской орфографии в XVII–XIX вв.: правописный узус и кодификация. Дисс. … докт. фил. наук. — М., 2010.

ВАРИАНТЫ РОДОВОЙ ПРИНАДЛЕЖНОСТИ В ГРУППЕ ИМЕН СУЩЕСТВИТЕЛЬНЫХ PLURALIA TANTUM В РУССКОМ ЯЗЫКЕ[1]

Савчук С. О. (savsvetlana@mail.ru)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Излагаются результаты корпусного исследования одного из участков вариативности морфологической системы русского языка. Результаты, полученные на материале разных корпусов в составе Национального корпуса русского языка, сравниваются с данными лексикографических источников и интернета. Анализируется соотношение вариантов, отмечается, что вариативность по роду представлена в группе pluralia tantum шире, чем в других группах существительных, и родовые варианты имеют более длительный период сосуществования и конкуренции.[2]

Ключевые слова: грамматическая вариативность, род существительных, корпусное исследование

В настоящей статье рассматривается состояние вариативности в группе существительных, характеризующихся нестандартным числовым поведением: во-первых, это существительные pluralia tantum, обладающие формальной дефектностью парадигмы, а во-вторых, потенциальные pluralia tantum, или существительные, преимущественно употребляющиеся во множественном числе. Эти существительные чаще привлекают внимание исследователей особенностями грамматической семантики[3]. Однако они представляют интерес и в плане изучения морфологической вариативности[4].

В силу особенностей семантики «истинные» pluralia tantum находятся вне категории родовых противопоставлений и образуют самостоятельный согласовательный класс — «парный род». Что касается потенциальных pluralia tantum, то многие из них допускают вариативное по роду образование форм единственного числа, при этом варианты по-разному оцениваются с позиции

[1] Работа выполнена при поддержке Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика» и РФФИ (грант 10-06-00151-а).

[2] Доклад публикуется в сокращенном варианте. Полная версия доступна на сайте конференции «Диалог».

[3] См. обзор литературы в Плунгян 2011, а также работу Ляшевская 2004.

[4] В нормативном аспекте проблема вариантов в группе потенциальных pluralia tantum рассматривается в Кадеева 1989, ГПРР; подробный анализ вариантов падежных форм существительных в том числе и этой группы см. в Обнорский 1931.

литературной нормы, что находит отражение в грамматиках и словарях правильностей. Помимо вариативности по роду группе pluralia tantum свойственна вариативность форм родительного падежа множественного числа, которая связана с родовой неопределенностью (*будни — буден* и *будней*, *бакенбарды — бакенбард* и *бакенбардов*). Таким образом, «истинные» pluralia tantum могут иметь только варианты род. п. мн. ч., а потенциальные могут присутствовать сразу в двух зонах вариативности. Кроме того, есть основания предполагать, что родовые варианты в группе pluralia tantum имеют более длительный период сосуществования и конкуренции, чем в других группах существительных, с течением времени определяющихся в выборе родовой соотнесенности и теряющих один из родовых коррелятов.

Исследование строится в рамках корпусного подхода и опирается на изучение состояния вариативности на большом массиве текстов, прочие лингвистические источники привлекаются только для сопоставления и интерпретации полученных результатов. Главное преимущество корпусного метода — это возможность получить достоверные и надежные сведения о частотности изучаемых явлений, поскольку для исследования используется замкнутый набор текстов. В частности, мы можем оценить соотношение вариантов и особенности их функционирования, опираясь на конкретные цифры, полученные на представительном материале. Другим преимуществом является возможность повторных исследований и верификации результатов.

Методика исследования

Методика исследования предполагает выполнение нескольких последовательных этапов.

1. Формирование первоначального словника лексем, вариативно образующих формы рода и/или формы родительного падежа множественного числа, для настоящего исследования проводилось на основе лексикографических источников[5]. Следующим этапом было расширение словника для обнаружения родовых и падежных вариантов у других слов группы, не учтенных в словарях и справочниках. С этой целью в состав словника были включены лексемы из Грамм, отнесенные к категории pluralia tantum (и имеющих соответствующую помету типа *<ж 1а>*) или потенциальных pluralia tantum (с пометой типа *мн. от ж 4а*). На основе предварительного отбора из списка были исключены лексемы, которые теоретически не могли иметь вариантов. Оставшийся

[5] Основные источники — словари начала XX в. (Чернышев 1915, Долопчев 1909, Огиенко 1914) и современные нормативные словари вариантов и трудностей (ГПРР, ТС, Еськова 1994). При этом важно подчеркнуть, что лексемы, представленные в словарях как имеющие вариативные формы только родительного падежа множественного числа, включались в список на общих основаниях с целью проверки наличия у них также и вариативных форм рода в единственном числе. Во многих случая прогноз оправдался, и вариативные родовые формы были отмечены у таких существительных, как босоножки, ласты, лампасы, литавры, корректив, онучи и др.

словник, объемом около 500 лексем, стал предметом исследования на материале НКРЯ[6].

2. Отбор и анализ контекстов употребления слова в корпусе текстов составляет содержание второго этапа работы. В нашем исследовании активное использование инструментария, предоставляемого НКРЯ (формирование запросов, средства поиска, автоматическая сортировка примеров, сохранение результатов поиска в среде Excel) неизбежно сочетаются с ручной морфологической разметкой словоформ, поскольку в русском языке в наиболее релевантных для диагностики значения числа и рода точках мы сталкиваемся с проблемой неоднозначности форм: *чулок* (род. мн = им. ед), *ботинки* (им. мн = род. ед. ж), *туфлю* (вин. ед. ж = дат. ед. м). Семантическая неоднозначность также снимается на этом этапе.

3. Сбор статистики и заполнение базы вариативных форм выполняется на третьем этапе. Выявленные на предыдущем этапе вариантные формы снабжаются информацией о грамматических и семантических признаках, количестве вхождений каждой словоформы и контекстах употребления. Кроме того, производится анализ словника и семантическая разметка лексем, на основе семантической классификации, принятой в НКРЯ. При этом омонимы и разные значения многозначных слов, если они не исключаются из рассмотрения на предыдущем этапе, получают разную интерпретацию.

4. Анализ и интерпретация полученных результатов могут быть грамматико-ориентированными, то есть описывать поведение целой группы вариантов, или лексикографически-ориентированными, представляющими характеристику отдельных лексем с их вариантными формами.

5. Практическим приложением работы является анализ корректности морфологической аннотации изучаемых словоформ, сбор информации об ошибках и исправление их, что ведет к совершенствованию лингвистической разметки в НКРЯ. В качестве другого практического выхода работы предполагается создание словаря вариантов, в котором актуальные для современного периода варианты были бы представлены в исторической перспективе [Савчук, Гришина, 2008].

Результаты исследования

В качестве примера использования методики корпусного анализа приведем фрагмент исследования вариантов в группе названий парных предметов.

---

[6] В перспективе результатом исследования вариативности стало бы пополнение состава вариантов за счет слов, извлеченных непосредственно из текстов корпуса. В частности, для рассматриваемой группы слов можно было бы отбирать леммы, которые представлены в текстах исключительно или преимущественно словоформами множественного числа. (См, например, реализацию этой идеи на материале BNC в ресурсе Sketch Engine (SkE) http://trac.sketchengine.co.uk/wiki/SkE/Histograms). Однако для русского языка эта задача осложнена проблемой снятия неоднозначности, требует лингвистического и программного обеспечения, и в настоящее время автоматически извлекать такого рода информацию из корпуса невозможно.

Значительное место в ней занимают названия обуви (всего было рассмотрено 37 лексем). Анализ форм, выражающих значения рода, показал, что вся группа по соотношению вариантов мужского и женского рода делится на четыре части:

(1) Существительные с вариативным способом образования форм единственного числа (таблица 1);
(2) Существительные, относящиеся в единственном числе к мужскому роду: *полуботинок, паголенок, катанок, мокасин, пим, чебот, бродень, торбас, ичиг, чувяк*, *котурн*.
(3) Существительные, относящиеся в единственном числе к женскому роду: *галоша, калоша, кроссовка, шиповка, крага, вьетнамка, балетка, сандалета, пинетка*.
(4) Существительные, у которых на материале НКРЯ не удалось определить значение рода: *боты, поршни, черевики, черевички, торбаза, галифаксы, бабуши, пьексы, баретки, карпетки, полукеды*.

В таблице 1 представлены слова с вариативными родовыми формами, ранжированные по соотношению вариантов мужского и женского рода.

Таблица 1

Варианты рода	муж	жен	всего вар.	ipm	%
валенок, м / валенка, ж	143	1	144	12,3	99
ботинок, м / ботинка, ж	1018	51	1069	27,56	95
опорок, м / опорка, ж	15	1	16	1,18	93
кед, м / кеда, ж	11	3	14	1,18	**79**
унт, м / унта, ж	5	2	7	0,78	**71**
штиблет, м / штиблета, ж	11	5	16	1,84	**69**
тапок, м / тапка, ж	24	19	43	1,67	**56**
ласт, м / ласта, ж	22	19	41	3,21	**54**
бахила, ж / бахил, м	2	2	4	0,52	**50**
тапочка, ж / тапочек, м	12	29	41	4,8	**29**
сандалия, ж / сандалий, м	14	48	62	3,58	**23**
босоножка, ж / босоножек, м	6	36	39	1,14	15
чуня, ж / чунь, м	1	11	12	0,45	8
онуча, ж / онуч, м	2	34	36	1,06	6
бутса, ж / бутс, м	1	18	19	0,58	5
туфля, ж / туфель, м	20	397	417	21,06	5

В начале списка расположены слова, в употреблении которых преобладают варианты мужского рода (*ботинок*, м / *ботинка,* ж), а в конце списка — слова с преобладающими формами женского рода (*туфля*, ж / *туфель,* м). В столбцах «муж», «жен» для каждой пары существительных приведено общее количество

форм мужского и женского рода, зафиксированных в основном корпусе НКРЯ, а в столбце «всего вар[иантов]» — их сумма. Столбец ipm показывает данные о частоте слова (как совокупности всех словоформ) в корпусе, и представляет собой отношение общего количества вхождений слова к объему корпуса, выраженному в миллионах словоупотреблений (в настоящее время — 209 млн). В последнем столбце показана доля варианта мужского рода.

Следует отметить, что в список, полученный на материале НКРЯ, вошли не только те слова, которые включены в известные словари вариантов (ГПРР, ТС — в таблице соответствующие строки выделены), но и такие пары, у которых варианты в словарях не фиксировались. Кроме того, относительно некоторых слов рекомендации справочников расходятся с данными корпуса. Так, в парах *бутс / бутса* и *кед / кеда* как более редкие варианты в ГПРР оцениваются *бутса* и *кед*, в то время как в текстах они лидируют с большим отрывом от своих корелятов. Варианты *унт* и *унта* оцениваются как равноправные, однако корпусные данные свидетельствуют об угасании варианта женского рода.

Соотношение вариантов

Если применить одну из шкал оценки соотношения варьирующих форм, предлагаемых исследователями, то **вариантными** должны быть признаны пары, у которых соотношение вариантов находится в интервале 84–16% (соответствующие им цифры в последнем столбце выделены жирным шрифтом). Формы м. р. *ботинок*, *опорок* и ж. р. *босоножка*, *туфля*, *чуня*, *онуча*, *бутса* считаются **доминантными** в своих парах (их доля составляет 85–100%), а в паре *валенок / валенка* вариант ж. р. рассматривается как **спорадический** (его доля меньше 2%).

Рассматриваемые варианты являют собой пеструю картину. С одной стороны, в группе много общеупотребительных слов с высокой частотностью: *ботинки, валенки, туфли, босоножки, сандалии, тапки* и *тапочки.* С другой стороны, в ней присутствуют слова с ограниченной сферой употребления, значительно менее частотные: либо относительно новые слова — обозначения обуви «специального назначения» (*кеды, бутсы, бахилы, ласты*), либо устаревшая лексика и регионализмы (*унты*, *чуни*), либо стилистически окрашенная, оценочная лексика (*штиблеты, опорки*).

С распределением вариантных пар внутри группы это расслоение лексики, как представляется, связано следующим образом. Самые частотные слова тяготеют к полюсам с доминированием вариантов мужского (*ботинок, валенок*) или женского (*туфля*) рода. В этих парах процесс конкуренции вариантов близок к завершению, говорящие прочно усвоили какую-то определенную модель словоизменения, которая закрепилась в парадигме, утратившей черты формальной дефектности[7].

[7] Это подтверждается анализом хронологии контекстов: в частотных словах *ботинок / ботинка, валенок / валенка* маргинальные формы женского рода зафиксированы

Среди вариантных пар, которые обнаруживают колебания в родовом соотнесении форм единственного числа, преобладают слова общеупотребительные. Можно предположить, что существование вариантов у таких слов имеет временный характер: у новых слов их конкуренция высока, но по мере освоения слова она завершается либо победой одного из вариантов, либо их семантическим расхождением. В ряде случаев определяющую роль в соотношении вариантов играет новизна не слова в целом, а именно форм единственного числа[8]. Колебания родовых форм иногда поддерживаются наличием производных к каждому из членов вариативной пары. Ср. *сандалия, сандалька, сандалья, сандалета* и *сандалий, сандалик, сандаль*.

Слова с ограниченной сферой употребления, не входящие в активный запас, как правило, не развивают в литературном языке вариативности, но могут сохранять ее, если она зафиксирована в словарях и справочниках. Оставаясь в низкочастотной зоне, слово сохраняет и свои грамматические характеристики, поскольку образование редких форм единственного числа вызывает затруднения и требует обращения к кодифицированным образцам[9].

Таким образом, соотношение вариантов внутри вариативной пары определенным образом зависит от прагматических факторов, в частности

в текстах XIX-го и начала XX-го века, а 3 употребления слова *ботинка* в текстах второй половины XX в. представляют собой случаи цитирования. Следовательно, эти слова определились в выборе родовой соотнесенности и с середины прошлого века ведут себя стабильно.

[8] Так, например, объяснение колебаний в родовой соотнесенности таких широко распространенных слов повседневной речи, как *тапки* и *тапочки,* по-видимому, следует искать в том, что формы единственного числа этих слов начали проникать в литературный язык совсем недавно. Первая фиксация в корпусе относится к началу 1940-х годов. *Жора потерял и вторую* ***тапочку****, он стоит теперь на сцене в одних серых носках* <…> [В. П. Беляев. Старая крепость (1937–1940)]. Таким образом, конкуренция вариантов в этой паре происходит на наших глазах, и исход ее еще не очевиден. К таким же молодым образованиям относятся и формы единственного числа слова *босоножки*. До середины XX в. все формы единственного числа, зафиксированные в корпусе, к босоножкам как обуви отношения не имеют, а являются наименованиями лиц женского пола: *Я-то женился сам на* ***босоножке****, только что на себе было, а ты вон капиталов требуешь.* [Д. Н. Мамин-Сибиряк. Хлеб (1895)]. Первое появление в корпусе формы единственного числа в значении обуви относится к 1957 г.: *Высокий носатый тракторист, стоявший спиной к кабинке, очень забавно дирижировал чьей-то* ***босоножкой***. [Борис Можаев. Саня]. В целом преобладают формы женского рода, но варианты мужского рода также встречаются: *«Бычок» падает в нескольких метрах от брюнетки, и она наступает на него каблуком своего* ***босоножка***. [Владимир Козлов. Sex and Violence (2002)].

[9] Обращает на себя внимание, что рекомендуемые словарями единственные формы опорок, онуча, чуня и др. совпадают с доминантными вариантами, выявленными на материале корпуса. Аналогичное распределение форм родительного падежа множественного числа в зависимости от принадлежности слова к активному и пассивному запасу было обнаружено в названиях военных [Савчук 2010], однако эта гипотеза нуждается в проверке на материале других групп.

от принадлежности слов к активному или пассивному запасу, что находит выражение в их частотных характеристиках[10].

Функционирование вариантов в газетном корпусе

Анализ вариантов в газетном корпусе проводился по той же методике, что и основное исследование, и имел целью сравнение результатов, полученных при обследовании функционирования одних и тех же слов в разных корпусах. По своим параметрам газетный корпус имеет сходство и различия с основным корпусом.

По своему объему (113 млн словоупотреблений) он почти равен корпусу современных (с 1950 г.) текстов (116 млн), но отличается от него большей стилистической и хронологической однородностью при тематическом и фактическом разнообразии, поскольку в нем собраны только тексты газет и информационных агентств за период 2000–2008 гг., а в основном корпусе представлены все сферы функционирования.

Названия бытовой лексики — не самый выгодный материал для исследования на материале газетных текстов, поскольку заранее можно ожидать, что большая часть таких слов будет иметь там невысокую частотность и употребляться преимущественно во множественном числе в однотипных конструкциях. Естественная среда обитания такой лексики — бытовая сфера, художественная литература, дневниковая проза и мемуаристика. Тем не менее для 18 слов из исследуемой группы все же удалось определить родовую принадлежность, часть результатов представлена в таблице 2.

Таблица 2

Варианты рода	муж	жен	всего вар	ipm	%
валенок, м	64	0	64	5,75	100
ласт, м	1	0	1	0,76	100
полуботинок, м	1	0	1	0,27	100
тапочка, ж / тапочек, м	6	8	14	5,6	43
тапка, ж / тапок, м	2	28	30	0,97	6,4
туфля, ж / туфель, м	1	21	22	8,8	4,5
босоножка, ж	0	7	7	11,6	0
кроссовка, ж	0	13	13	6,72	0
сандалия, ж	0	11	11	1,42	0
калоша, ж	0	55	55	1,2	0

[10] Ср. с понятием освоенности слова как прагматическом факторе, влияющем на эволюцию акцентной системы (Зализняк 1985, 12).

Зону вариативности составляют 3 слова — *туфля / туфель, тапка / тапок* (с доминантными вариантами женского рода) и *тапочка / тапочек* с почти равноправными формами женского и мужского рода). Остальные слова, для которых удалось определить родовую принадлежность, образуют формы единственного числа в строгом соответствии с рекомендациями словарей. Часть слов оформляет варианты мужского рода: *ботинок, валенок, ласт, полуботинок, ботик, бредень, чувяк* и другая часть — варианты женского рода: *босоножка, кроссовка, сандалия, калоша, галоша, шлепка, шиповка, балетка, онуча*.

Таким образом, основной вывод, который можно сделать на основе сравнения данных, полученных при обследовании разных корпусов, состоит в следующем. Результаты, полученные на газетном корпусе, не противоречат результатам, полученным на основном материале, а отличия объясняются спецификой текстов. Газетные тексты более последовательно отражают действующий стандарт литературного языка, но не дают представления о его богатой вариативности. Вопреки расхожему мнению об общем снижении стиля прессы, журналисты и редакторы газетных издательств соблюдают кодифицированную литературную норму в области морфологии и строго ей следуют. В этом их позиция отличается от установки редакторов книжных издательств, предоставляющих авторам большую индивидуальную свободу в выборе вариативных грамматических средств.

Варианты рода по данным интернета

Большая часть слов рассматриваемой группы, в частности специализированных наименований обуви, получила широкое распространение в повседневной речи. В разговорной речи эти слова становятся частотными, что повышает вероятность употребления их в контекстах, требующих образования форм единственного числа, а значит, и возможность реализовать родовую вариативность без оглядки на литературный стандарт. Употребление этих слов в бытовой сфере характеризуется отклонениями от литературной нормы. Богатый материал предоставляет интернет; в блогах и рекламных текстах можно обнаружить вариантные пары практически ко всем литературным формам[11]. **Кроссовка — кроссовок**, м: *У меня при беге этот шов на левом кроссовке давит на верхнюю часть ступни, на тренировке по мере увеличения дистанции дискомфорт проходит. С правым кроссовком такого не происходит*. **Шиповка — шиповок**, м: P. S. *Мне самому интересно, что же стало с тем человеком, которому*

[11] Количественно преобладают, конечно, нормативные варианты. В приводимых в качестве примеров парах иллюстрируются нестандартные варианты. Интересно, что эти же отличия от литературно-грамматической нормы в употреблении форм рода и числа существительных называет В. В. Виноградов среди морфологических особенностей «низового» городского говора середины XIX в., при этом замечает, что многие явления, первоначально относившиеся к городскому просторечию, постепенно вошли в систему литературного языка [Виноградов 1938, 480, 481].

отдавил ногу шиповком… **Пинетка — пинеток**, м: *На фото с одним пинетком — чтоб было понятно на сколько пинеток увеличивает шаговый шов. В комплекте конечно же оба. по всем вопросам пишите смс*. **Вьетнамка — вьетнамок**, м: *Мне подруга сегодня привезла из Германии пару стирачек в форме тапков-вьетнамков….знает мои проблемы))) Я с досадой пинаю вьетнамком песок и иду к бару…* **Полукеды — полукед**, м: *На чудо-дереве растут кеды и полукеды. Если сорвать один кед или полукед, то он вырастет снова*.

Как представляется, сфера разговорной речи в ее устной и письменной электронной форме, городское просторечие, различные жаргоны и под. в настоящее время поставляют большое количество субстандартных вариантов (таких как рассмотренные выше родовые варианты или склоняемые формы, соответствующие несклоняемым существительным литературного языка[12]), часть которых со временем может быть кодифицирована.

Выводы

Как показало проведенное исследование, вариативность по роду характерна для современного состояния русского литературного языка в гораздо большей степени, чем это отражено в словарях вариантов и грамматических описаниях, и широко представлена прежде всего в группе существительных pluralia tantum, а также в группе несклоняемых существительных.

Главным системным фактором, определяющим возникновение и существование вариантов в группе pluralia tantum, является отсутствие строгой предсказуемости родовой формы единственного числа на основании формы множественного числа и как следствие возможность разных способов выражения одного и того же грамматического значения. Вследствие этого вариативные пары имеют более длительный период сосуществования и конкуренции вариантов.

Семантический фактор определяет специфику взаимоотношений вариантов в разных семантических группах разных уровней иерархии, например в группах предметных и непредметных имен, группах названий парных и сложных предметов, совокупностей, веществ, действий и т. д. Общие выводы относительно тенденций развития вариативности во всей группе существительных можно делать на основании наблюдений за поведением вариантов во всех выделяемых в ней подгруппах.

Важным фактором, определяющим продолжительность существования вариантов, является частота употребления слова, а также отдельных его форм. В общем случае чем выше частотность слова, тем быстрее оно осваивается устной и письменной речью и тем быстрее за ним закрепляется какая-то одна родовая характеристика.

[12] О результатах исследования родовой вариативности в группе несклоняемых существительных на материале НКРЯ см. Савчук 2011.

Сохранению вариантных пар способствует группа факторов, которые можно отнести к стилистическим. Во-первых, это закрепление разных вариантов в разных сферах общения, в литературном языке и субстандарте. Во-вторых, это функциональное расслоение вариантов, когда за одним из них закрепляется какая-либо стилистическая окраска (*прост.*, *устар.*, *спец.*), которая может широко использоваться для целей художественной изобразительности.

Возможные области практического применения полученных данных — коррекция морфологической разметки НКРЯ, а также лексикографическое описание вариантов в диахронической перспективе.

Литература

1. *Виноградов* 1982 — Виноградов В. В. Очерки по истории русского литературного языка XVII — XIX веков. Изд. 3-е. М.: Высшая школа, 1982.
2. *ГПРР* — Граудина Л. К., Ицкович В. А., Катлинская Л. П. Грамматическая правильность русской речи. М.: Наука, 1976; Изд. 3-е: 2004.
3. *Грамм* — Зализняк А. А. Грамматический словарь русского языка. Изд. 4-е, испр. и доп. М.: Русские словари, 2003.
4. *Долопчев* 1909 — Долопчев В. Опыт словаря неправильностей в русской разговорной речи. Изд. 2-е. Варшава, 1909.
5. *Еськова* 1994 — Еськова Н. А. Краткий словарь трудностей русского языка. Грамматические формы. Ударение. М.: Русский язык, 1994.
6. *Зализняк* 1985 — Зализняк А. А. От праславянской акцентуации к русской. М.: Наука, 1985
7. *Кадеева* 1989 — Кадеева М. И. Вариативность в группе потенциальных pluralia tantum (нормативный аспект). АКД … канд. филол. н. М., 1989.
8. *Ляшевская* 2004 — Ляшевская О. Н. Семантика русского числа. М.: Яз. славян. культур, 2004.
9. *Обнорский* 1931 — Обнорский С. П. Именное склонение в современном русском языке. Изд. 2-е (1931) — М.: Книжный дом «Либроком», 2010.
10. *Огиенко* 1914 — Огиенко И. И. Словарь неправильных, трудных и сомнительных слов, синонимов и выражений в русской речи. Изд. 3-е. Киев, 1914.
11. *Плунгян* 2011 — Плунгян В. А. Введение в грамматическую семантику: грамматические значения и грамматические системы языков мира. М., 2011
12. *Савчук, Гришина* 2008 — Савчук С. О., Гришина Е. А. Вариантность в русском языке. Проект словаря // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (2008). Вып.7. С. 466–474.
13. *Савчук* 2010 — Савчук С. О. Опыт корпусного исследования морфологической вариативности: варианты родительного падежа множественного числа существительных мужского рода // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (2010). Вып. 9 (16). С. 419–429.

14. *Савчук* 2011 — Савчук С. О. Корпусное исследование вариантов родовой принадлежности имен существительных в русском языке // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (2011). Вып. 10. С. 562–579.
15. *ТС* — Трудности словоупотребления и варианты норм русского литературного языка: Словарь-справочник / Под ред. К. С. Горбачевича. Л.: Наука, 1973.
16. *Чернышев* 1915 — Чернышев В. Правильность и чистота русской речи: Опыт русской стилистической грамматики. // В. И. Чернышев. Избранные труды. Т. 1. М.: Просвещение, 1970. С. 443–641.

TOWARDS THE AUTOMATED ENRICHMENT OF MULTILINGUAL TERMINOLOGY DATABASES WITH KNOWLEDGE-RICH CONTEXTS

Schumann A.-K. (anne.schumann@tilde.lv)

Tilde SIA/University of Vienna, Riga/Vienna, Latvia/Austria

This paper describes ongoing Phd thesis work dealing with the extraction of knowledge-rich contexts (KRCs) from specialized Russian and German text corpora for the semantic enrichment of terminological resources. In recent years, automatically extracted KRCs have been proposed as a means for deriving empirically grounded concept descriptions for terminography while maintaining the time and costs spent for the acquisition of such descriptions on a reasonable level. KRCs have been studied for a number of European languages ranging from English over French and Spanish to Catalan, however, not much effort has yet been put into widely spoken, but typologically different languages such as Russian or German. This paper, therefore, describes research efforts aiming at the extraction of KRCs in Russian and German for the purpose of termbase enrichment. Section 1 of this paper presents a brief introduction to KRC research and the motivation for this study. Section 2 gives an overview over related work. Section 3 describes the KnowPipe KRC extraction framework, whereas section 4 outlines ranking experiments with KnowPipe on Russian and German data. Section 5 summarizes the results and describes future work.

Key words: Corpus linguistics, knowledge-rich contexts, computer-aided terminography, Russian language, German language

1. Introduction

Definitions and explanations of concepts are an obligatory part of any termbase entry (ISO, 2009). However, there is no framework for the systematic enrichment of termbases with such content. In practice, semantic information is often added manually and unsystematically or omitted completely because of practical constraints. In this context, knowledge-rich context (KRC) extraction aims at identifying *semantic contexts* (as opposed to *linguistic contexts*) that provide *semantic information* about *concepts* (as opposed to *linguistic information* about *terms*) in text corpora and to feed the results of this process into a terminological resource. In other words, KRC extraction aims not only at providing examples of term occurrences, but sentences that provide additional knowledge about the concept to which a term refers. Research in this field, therefore, touches on aspects of terminology research that remain yet unresolved: Although the semantic types of contexts have been described in ISO 12620 (ISO, 2009), many terminological resources do not distinguish between different types of contexts and mainly restrict themselves to linguistic contexts and more or less informative usage examples. Often, however, contexts are completely omitted.

The aim of this study is to investigate, to which extent KRC extraction can contribute to the enrichment of a terminological resource with additional information by describing experiments conducted on two Russian corpora as well as on one German language corpus. In our experiments, the internet is used as a source of information since it is a primary means for finding information about terms and concepts for many professional translators and interpreters, and, in our view, a KRC extraction approach must therefore be able to deal with the quality of data found online in order to be applicable to real tasks.

2. Related Work

KRC extraction generally requires high precision, while specialized corpora from which KRCs can be extracted are typically small or must be crawled from online sources, a process that often outputs messy data. What is common to most studies in the field, therefore, is the fact that they employ a pattern-based method for KRC extraction. A systematic overview over pattern-based work is given by Auger and Barrière (2008). In most approaches, extraction patterns are acquired manually, but some groups (Condamines & Rebeyrolle, 2001; Halskov & Barrière, 2008) also devise a bootstrapping procedure for automated pattern acquisition similar to methods developed in information extraction (Xu, 2007). Seminal work for English was carried out by Pearson (1998) and Meyer (2001), and more recent work providing a contrastive linguistics perspective on English and French is Marshman (2007) and Marshman (2008). Recent studies for other languages are Feliu & Cabré (2002) for Catalan, Sierra et al. (2008) for Spanish, and Malaisé et al. (2005) for French. Sierra et al. (2008) employ 13 verbal patterns to extract instantiations of 3 previously defined types of definitions and implement a set of rules for filtering out irrelevant candidates. Malaisé et al. (2005) use a total of 42 verbal patterns for French as well as metalinguistic markers such as parentheses. For German, KRC extraction has not been studied, but Walter (2010) provides a detailed account on the related topic of extracting definitions from court decisions. A recent study (Schumann, 2011) compares KRC extraction from German and Russian corpora.

As for the ranking of extraction output, Walter (2010) gives a detailed account of his experiments in the ranking of definition candidates using supervised machine learning techniques. The features used in his experiments can be divided into five groups:

- *Lexical*, such as boost words or stop words and features that are specific for legal language, such as subsumption signals
- *Referential*, such as anaphoric reference or definiteness of the definiendum
- *Structural*, such as the position of the definiendum relative to the definiens
- *Document-related*, such as the position of the definition candidate in the document and whether there are other candidates in its immediate context
- *Others*, such as sentence length or TF-IDF scores of terms in the sentence

Walter produces the best results using a linear regression algorithm. He also carries out experiments using the output of supervised classifiers such as Naïve Bayes or k-Nearest Neighbour as an additional feature in ranking.

3. Knowledge-Rich Context Extraction in Russian and German

3.1. Extraction Patterns

Previous studies of KRC extraction from Russian and German web corpora (Schumann, 2011) were based on a pattern-based extraction approach using mainly predicative Russian and German patterns. These patterns were combined either with target terms or morpho-syntactic term formation patterns to form regular expressions. Example 1 illustrates a lexical extraction trigger used in our Russian experiments and a valid KRC. The underlined term is the target term, the lexical extraction trigger is marked in bold. Example 2 illustrates a German KRC.

(1) ***Система охлаждения*** *служит для отвода излишнего тепла от деталей двигателя, нагревающихся при вго работе.*
[Translation: The ***cooling system*** *serves to remove excess heat from those parts of the engine that heap up during exploitation.]*

(2) *Das Verhältnis Energieertrag (Output) zu Input wird* ***Leistungszahl*** *genannt.*
[Translation: The relation between energy output and input is called ***coefficient of performance****.]*

Example 2 also illustrates that in KRC extraction from German text, it is necessary to deal with considerable syntactic complexity, since predicates used as lexical extraction triggers often consist of several surface words that are distributed across the sentence forming long syntactic dependencies, into which the target term can be embedded — a problem that deserves further investigation.

Table 1 gives a simplified overview over the lexical patterns currently used by our extraction framework KnowPipe and assigns them to earlier defined semantic target relations (Schumann, 2011):

Table 1. Semantic relations and Russian extraction triggers

Relation	Explanation	Russian Patterns	German Patterns
Hyperonymy	Generic-Specific	Относить к, относиться к, включать в себя, классифицировать, различать, подразделять, разделять на, разделяться на, входить, составлять:, различать следующий, различать ... тип	Gehören zu, eine Form von ... sein, sein unter

Relation	Explanation	Russian Patterns	German Patterns
Meronymy	Part-Whole	Состоять из, включать в себя, снабжать, снабдить, образовать, составлять:, оснащать, оснащаться, оснащенный	Bestehen aus, (sein|werden) (versehen | ausgestattet | ausgerüstet | bestückt), besitzen, verfügen über
Process	Temporal neighbourhood	Воздействовать, приводить, осуществлять	bewirken
Position	Spatial neighbourhood	Расположенный, располагать, устанавливать, устанавливаться	Sein (positioniert | angeordnet)
Causality	Cause-Effect	Обусловить, обуславливать, обеспечить, определяться	Ergeben sich aus, bewirken
Origin	Material or ideal origin	Состоять из	–
Reference	General predication or definition	Представлять себя, называть, называться, определить, понимать, -это, так.называемый	(sein | werden) durch ... (charakterisiert | gekennzeichnet | beschrieben), werden genannt, (nennt | bezeichnet | unterscheidet) man, werden als ... bezeichnet, unter versteht man, werden (über | durch | mit) beschrieben, stellen dar, so.genannt
Function	Purpose or aim	Служить, позволять, предназначать, нужный для, применять, применяться	Stellen sich die Aufgabe, haben die (Aufgabe | Funktion), dienen (dazu|als), zu (nutzen | einsetzen | verwenden | benutzen)

The differences between the German and Russian pattern lists suggest that the latter are not yet definite, but need to be enlarged as soon as more information becomes available, e. g. about less-researched relations such as Origin.

3.2. KnowPipe

KnowPipe aims at providing a processing environment for multilingual KRC extraction using shallow as well as deep processing. In its current state of development,

KnowPipe offers preprocessing tools for Russian and German text corpora, pattern-based KRC extraction as well as a ranking method based on shallow features for both languages.

Preprocessing consists of tokenization, removal of duplicate sentences, removal of stop sentences (e. g. incomplete sentences, questions) and lemmatization. The Perl Lingua::Sentence module[1] is used for tokenizing Russian and German corpora. TreeTagger (Schmid, 1994) is used for lemmatization. KRC candidates are extracted using the lexical patterns described in the previous section. They are then ranked directly according to the values outputted by a supervised machine learning algorithm, currently Naïve Bayes. The Naïve Bayes algorithm was chosen since it seemed to generalize best to different types of input data. The Perl Algorithm::NaiveBayes module[2] is used to carry out this process. It uses the following 13 features for ranking KRC candidates both in Russian and German:

Table 2. Shallow features used for ranking

Feature name	Explanation
Word tokens	The number of word tokens in the sentence.
Subscore	The normalized sum of the term relevance scores of terms constituting the subject.
Subpos	1 if the sentence starts with the subject, else 0.
Term score	The normalized sum of the term relevance scores scores of all other terms[3].
Nr. of terms	The number of terms in the sentence.
Position	1 if the subject is located before the extraction pattern, else 0.
Adjacent term	1 if there is a term directly adjacent to the extraction pattern, else 0.
Distance	An integer indicating the token distance between subject and pattern.
Negation	1 if the extraction pattern is preceded by a negation particle, else 0.
Boost words	1 if the pattern is preceded by a generalization signal, else 0.
Pattern score	A pattern reliability score based on baseline experiments (Schumann, 2011).
Stop words	Number of negative markers normalized by word tokens.
Definite Subject	1 if the subject is preceded by markers of definiteness or anaphora, else 0.

[1] http://search.cpan.org/~achimru/Lingua-Sentence-1.00/lib/Lingua/Sentence.pm.

[2] http://search.cpan.org/~kwilliams/Algorithm-NaiveBayes-0.04/lib/Algorithm/NaiveBayes.pm.

[3] An appropriate subsection of the Russian Internet Corpus (Sharoff, 2006) was used as a reference corpus in scoring. A search interface to this corpus is available here: http://corpus.leeds.ac.uk/ruscorpora.html. For German, we used a subsection of a 2011 newscrawl from the Leipzig corpora collection (Quasthoff et al., 2006), http://corpora.informatik.uni-leipzig.de/download.html.

For Russian, we devised a heuristic that uses the rich annotation provided by the Russian TreeTagger tagset (Sharoff et al., 2008) and syntactic noun phrase formation patterns to identify noun phrases in nominative case. For German, we used the Parzu dependency parser (Sennrich et al., 2009) for analyzing sentence structure. Boost words are expressions such as *обычно, чаще всего* etc. that are indicative of a generalized statement for Russian and *üblicherweise, oft, häufig* etc. for German. Stop words, on the other hand, are expressions that are usually used for creating coherence, especially when the information provided in the sentence is single-case information and therefore not generalizable or makes reference to information that was given earlier in the text. Examples for stop words are: *иными словами, в итоге, поэтому, кроме того, например* etc for Russian and *dabei, aus diesem Grund, aber* etc. for German. The positional features are based on the hypothesis that even in a free word order language like Russian or German, KRCs favour canonical syntax over inverted structures. Moreover, a definite subject is treated as an indicator for single-case information which usually does not form a part of KRCs.

4. Experiments

4.1. Corpora and Experimental Setup

The performance of pattern-based KRC extraction from Russian and German text corpora has already been studied in Schumann (2011). For evaluating the performance of the ranking algorithm, we conducted experiments on the data outputted by this extraction step. For Russian, this data was extracted from two corpora, namely a small corpus dealing with the automotive domain and a larger corpus covering several engineering topics, as described in Schumann (2011). For German, KRC candidates extracted from a rather large, but noisy corpus covering the domains of electrical engineering and wind energy were used. For all corpora, a gold standard had been created earlier by means of manual annotation.

370 KRC candidates from the Russian automotive corpus and 709 KRC candidates from the larger Russian corpus were used for ranking. Since the ranking is carried out by means of a supervised learning algorithm, each data set had to be split into a training set in which each KRC candidate is marked as valid or invalid KRC and a test set on which the actual ranking is performed. On the Russian car corpus, 100 sentences were used for training and 270 for testing. On the Russian multidomain corpus, this relation was 300/409. The training sets were kept small to ensure the usability of the algorithm in a practical extraction task where typically not much data can be annotated manually. 322 frequently occurring terms were manually extracted from the gold standard for the car corpus. For the Russian multidomain corpus, the corresponding number was 372. These terms served as target terms — terms for which the extracted KRCs may supply additional information — in the feature annotation step. For German, the sample of KRC candidates extracted from the German multidomain corpus comprised 574 sentences. Out of these, 200 were manually annotated for

training, the rest was used for testing. As for the Russian data sets, the earlier created gold standard was used as a reference. However, the number of manually extracted target terms was much higher for German than for Russian, namely 526.

4.2. Evaluation

For all three test samples Precision before and after ranking was calculated using the respective gold standards as reference. Table 3 gives an overview over the Precision values achieved on both Russian corpora as well as the German corpus before and after ranking for different Recall levels. Note that Recall was calculated with respect to the whole corpus, not just the test sample.

Table 3. Precision for different Recall levels before and after ranking

Recall levels	0,10	0,20	0,30	0,40
Russian car corpus				
Precision before ranking	0.49	0.45	0.47	0.51
Precision after ranking	0.97	0.85	0.73	0.49
Russian multidomain corpus				
Precision before ranking	0.38	0.40	0.35	–
Precision after ranking	0.89	0.66	0.38	–
German multidomain corpus				
Precision before ranking	0.47	0.4	0.33	–
Precision after ranking	0.8	0.69	0.45	–

5. Discussion and Future Work

The results outlined in the previous section are encouraging in the sense that the ranking algorithm seems to support the selection of valid KRCs: Valid KRC candidates are moved to the top of the sample, whereas invalid candidates are moved to the bottom, which produces relatively high Precision in the top-n segment and lower Precision for the bottom of the test sample, allowing for an easier selection of valid KRCs at the end of the process. Since no language-specific features are used, the ranking algorithm also generalizes to German. However, the results obtained on the German data still need to be considered preliminary as no experiments with a second corpus have been conducted yet. In terms of overall performance — especially with respect to a real-world database enrichment task — several improvements still need to be undertaken, including the improvement of overall Precision which is likely to result in a better performance also during ranking. In the future, we plan to conduct experiments on using deeper linguistic knowledge, e. g. syntactic information, for creating more powerful — and thus more precise — extraction patterns. This seems especially relevant to the case of German, however, we believe that the use of more linguistic

information will improve the performance of both the extraction step and the ranking algorithm also on the Russian data. Recall still calls for further improvement and experiments need to be conducted on larger datasets. For termbase enrichment, means need to be devised for defining the target term of each KRC candidate, e. g. for establishing an explicit relation between an extraction pattern and a term that is being explained by the phrases introduced by the pattern. There is reason to believe that richer linguistic information will also be helpful in this task.

Acknowledgement

The research described in this paper was funded by the CLARA project (EU FP7 /2007–2013), grant agreement n° 238405.

References

1. *Auger, A., Barrière, C.* (2008), Pattern-based approaches to semantic relation extraction. Terminology, Vol. 14 (1), pp. 1–19.
2. *Condamines, A., Rebeyrolle, J.* (2001), Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB). In D. Bourigault, C. Jacquemin, M.-C. L'Homme (Eds.), Recent Advances in Computational Terminology, Amsterdam/Philadelphia, John Benjamins, pp. 127–148.
3. *Feliu, J., Cabré, M.* Conceptual relations in specialized texts: new typology and an extraction system proposal. Proceedings of TKE, Nancy, 2002, pp. 45–49.
4. *Halskov, J., Barrière, C.* (2008), Web-based extraction of semantic relation instances for terminology work. Terminology, Vol. 14 (1), pp. 20–44.
5. *International* Organization for Standardization (2009), International Standard ISO 12620: 2009 –Terminology and Other Language and Content Resources – Specification of Data Categories and Management of a Data Category Registry for Language Resources. Geneva, ISO.
6. *Malaisé, V., Zweigenbaum, P., Bachimont, B.* (2005), Mining defining contexts to help structuring differential ontologies. Terminology, Vol. 11 (1), pp. 21–53.
7. *Marshman, E.* (2007), Towards strategies for processing relationsips between multiple relation participants in knowledge patterns. An analysis in English and French. Terminology, Vol. 13 (1), pp. 1–34.
8. *Marshman, E.* (2008), Expressions of uncertainty in candidate knowledge-rich contexts. A comparison in English and French specialized texts. Terminology, Vol. 14 (1), pp. 124–151.
9. *Meyer, I.* (2001), Extracting Knowledge-Rich Contexts for Terminography: A conceptual and methodological framework. In Recent Advances in Computational Terminology, pp. 279–302.
10. *Pearson, J.* (1998), Terms in Context. (Studies in Corpus Linguistics 1). Amsterdam/Philadelphia, John Benjamins.

11. *Quasthoff, U., Richter, M., Biemann, C.,* Corpus Portal for Search in Monolingual Corpora, Proceedings of LREC, Genoa, 2006, pp. 1799–1802.
12. *Schmid, H.,* Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of the International Conference on New Methods in Language Processing, Manchester, 1994, pp. 44–49.
13. *Schumann, A.-K.,* A Bilingual Study of Knowledge-Rich Context Extraction in Russian and German. Proceedings of the Fifth Language & Technology Conference, Poznan, 2011, pp. 516–520.
14. *Sennrich, R., Schneider, G., Volk, M., Warin, M.,* A New Hybrid Dependency Parser for German. Proceedings of GSCL Conference, Potsdam, 2008.
15. *Sharoff, S.* (2006), Creating general-purpose corpora using automated search engine queries. In Baroni, M., Bernardini, S. (Eds.), WaCky! Working papers on the Web as Corpus. Bologna, Gedit.
16. *Sharoff, S., Kopotev, M., Erjavec, T., Feldmann, A., Divjak, S.,* Designing and evaluating Russian tagsets. Proceedings of LREC, Marrakech, 2008.
17. *Sierra, G., Alarcón, R., Aguilar, C., Bach, C.* (2008), Definitional verbal patterns for semantic relation extraction. Terminology, Vol. 14 (1), pp. 74–98.
18. *Walter, S.* (2010), Definitionsextraktion aus Urteilstexten. PhD thesis in Computational Linguistics. Saarland University Saarbrücken.
19. *Xu, F.-Y.* (2007), Bootstrapping Relation Extraction from Semantic Seeds. PhD thesis in Computational Linguistics. Saarland University Saarbrücken.

РУССКОЕ ИМЯ ПАРАМЕТРА: МЕТАФОРИЧЕСКИЕ И МЕТОНИМИЧЕСКИЕ ПРОЦЕССЫ

Семенова С. Ю. (Sonya_sem@mail.ru)

Институт научной информации по общественным наукам РАН, Российский государственный гуманитарный университет, Москва, Россия

Рассматриваются производные (метафорические и метонимические) значения и производные таксономические категории участника для русских параметрических существительных. Метафоризация наименования исходной физической величины приводит к семантическому сдвигу параметрического имени либо с сохранением, либо с утратой измеримости референта. Соответственно, вводятся понятия слабой и сильной метафоризации. Изменение таксономии участника также бывает обусловлено метонимическими сдвигами в нем; возможные классы участника могут быть генетически упорядочены в словаре. У ряда параметрических имен отмечается также собственный метонимический переход - к предметному значению. Исследование, с одной стороны, имеет теоретическую направленность - на получение детальных когнитивных схем семантической деривации; с другой стороны, оно нацелено на прикладное словарное описание параметрической лексики для задач автоматического извлечения информации.

Ключевые слова: параметрическое существительное, метафоризация, метонимический перенос, семантика актанта

1. Введение. Исходная полисемия параметрического имени

Целью работы является рассмотрение определенных типов полисемии русских параметрических существительных, а именно, дериватов от обозначения исходного количественного параметра (ср.: *высота мачты –> высота звука*, *высота помыслов* и др.). Индикатором семантического переноса во многих случаях является смена семантического (таксономического) класса первого актанта параметрического имени, т. е. объекта, характеризуемого по параметру. На взаимосвязь значения предикатной лексемы и таксономии ее актантов указывается в работах Е. В. Падучевой [Падучева 2002 и др.], где таксономические классы участников интерпретируются как параметры лексического значения. Для собственно параметрических имен таксономия участника рассматривается в [Семенова 2008]; наст. работа написана в развитие последней.

Полисемия весьма характерна для параметрических имен; так, в литературе неоднократно указывалось на регулярную полисемию ‘параметр — большое

значение параметра' (ср.: *вычислить высоту* и *бояться высоты*). В первую очередь это свойство отмечалось для имен линейных размеров (*высота, длина*, *ширина*, *толщина*, *глубина* [Апресян 1992 и др.]), хотя оно справедливо и для существенно более широкого круга параметров: ср.: *измерить температуру* и *у больного температура* (во втором случае температура повышенная); *пенсия по возрасту* (=> возраст солидный); *количество переходит в качество* (=> количество достаточно большое); *работать в темпе* (=> темп быстрый) и т. п.

Имена параметров, относясь к категории абстрактной, т. е. прототипически более «поздней» лексики (по сравнению с лексикой конкретной, в том числе предметной), во многих случаях тоже являются семантическими дериватами (т. е. не исходными значениями данного слова), напр., *радиус*-параметр есть дериват от *радиуса*-фигуры, *сопротивление* как физическая величина производно от общелексического *сопротивления*, медицинский параметр *пульс* (частота сердечных сокращений) — дериват от названия явления биения сердца, *квадрат* (математическая функция) есть результат абстрагирования от названия фигуры и т. п. Собственно, исконные отадъективные имена параметров, в ряде отношений ядерные для данного класса (*высота*, *глубина*, *частота*, *мощность* и др.), также представляются производными от субстантивировавшихся наименований свойств со значением верхнего полюса (*высота*-параметр естественным образом мыслится в ретроспекции как дериват от *высоты*-свойства, т. е. от большой высоты). Так, Е. Курилович выделял два этапа в процессе образования имен типа *высота*: этап синтаксической деривации, на котором имеет место субстантивация прилагательного *высокий*, выражающего свойство «быть высоким», и этап лексической деривации, на котором у субстантива возникает значение параметра [Курилович 1962, с. 64–65].

В данной работе нас будет интересовать дальнейшая семантическая деривация имени параметра — от исходного (например, пространственного) параметрического количественного значения (с исходным участником, например, предметным) к употреблениям производным и, в частности, более абстрактным. В соответствующих деривационных процессах может иметь место сохранение (или наоборот, утрата) измеримости, сохранение или утрата параметричности в широком смысле (т. е. обозначения варьируемой характеристики, см. ниже); наблюдаются и трансформации в участнике — измеряемом объекте. Изучение соответствующих процессов, выявление регулярных моделей преобразований, как представляется, ложится в русло когнитивного подхода, для которого параметрическая лексика традиционно служила материалом исследований, полигоном при формировании и апробации методов.

Изменения семантики параметрических имен, в т. ч. утрата измеримости и мена физической величины, существенны и для прикладных задач, в первую очередь для автоматического извлечения параметрической информации. Если параметрическое имя, встреченное в тексте, не обозначает измеряемого параметра, то алгоритм должен распознать и проигнорировать такое употребление. Если изменилась природа измеряемой величины, это также должно быть распознано. Если вследствие трансформаций меняется класс участника, полезно

выявить исходный класс. Все указанные подзадачи должны опираться на релевантное грамматическое и словарное описание.

Таким образом, исследование приобретает двоякую направленность. С одной стороны, оно нацелено на выявление когнитивной «картины» семантических преобразований, в т. ч. отмеченных регулярностью (и на этом пути сделаны лишь начальные шаги), с другой стороны, на уточнение прикладных описаний.

2. Последующие деривационные процессы

При рассмотрении семантических процессов будем опираться на генитивную конструкцию с параметрическим именем в вершине (*глубина реки*) — такая конструкция обладает наиболее широкой сочетаемостью среди атрибутивных конструкций с именами параметров, она индифферентна к наличию/отсутствию измеримости и, и соответственно, независима от второго актанта параметрического имени — количественной группы (ср., например, данную конструкцию, где количественное значение синтаксически факультативно: *высота строения*, с другой распространенной атрибутивной конструкцией, с инструментативом: *строение высотой 7 метров*, где количественное значение облигаторно).

Итак, в генитивных группах с параметрическим именем, при деривации от исходного параметрического употребления наблюдаются следующие семантические процессы (сопровождающиеся изменениями в участнике).

1. Метафоризация параметрического имени с сохранением параметрического количественного значения: *глубина паза, глубина сверления, глубина резки металла, глубина просмотра гипертекста (или сайта); плотность тока, плотность населения, плотность дефектов кристаллических структур* и т. п. В этих примерах у имени *глубина* утрачивается исходный компонент ‘дно’, ‘направленность вертикально вниз’, но сохраняется идея ‘направленность внутрь некой среды’; ср. толкование исходного параметрического значения имени *глубина*: «глубина — протяженность, расстояние от поверхности до дна или до какой-нибудь точки по направлению вниз» [Словарь С. И. Ожегова]. (Кроме того, в сочетаниях *глубина сверления, глубина резки металла* отмечается метонимия в участнике, см. ниже п.3.) У имени *плотность* утрачивается исходный компонент ‘масса вещества’, но сохраняется идея ‘количество некой субстанции / в единице пространства’. При этом у обоих имен сохраняется параметрическое количественное значение — данные *глубины* и *плотности* подлежат измерению. Такую метафоризацию, преобразующую лексическое значение, но не выводящую слово за сферу количественной параметрики, назовем «слабой». Слабая метафоризация наблюдается у ряда параметрических имен (*высота*, *емкость*, *объем*, *мощность* и нек. др.): *высота тона*, *высота небесного светила*, *емкость конденсатора*, *объем капитальных вложений*, *объем памяти* (компьютера), *мощность множества* и т. п.

Приведенные имена параметров отличаются определенной образностью, «наглядностью», что представляется существенным условием для развития переносных значений (в отличие от имен величин, являющихся исходно более «специальными», таких напр., как *проводимость* или *теплоемкость*). Наименования величин, которые развивают (по крайней мере, в первую очередь) метафорические значения, должны быть «понятны», естественны для обычной картины мира. Именно к ним, понятным и наглядным, нередко прибегают исследователи для обозначения новых понятий в предметных областях (по аналогии, опираясь на обыденный опыт). Кроме того, в исходных значениях метафоризирующиеся имена относительно конкретны, «физичны» (в отличие, например, от общепараметрических имен типа *количество*, *коэффициент*, которые не метафоризуются в силу изначально высокой степени абстракции).

В ряде случаев косвенным показателем трансформации лексического значения выступает смена единицы измерения (*плотность тока* измеряется в иных единицах, нежели исходная *плотность*; математическая величина *мощность множества* в отличие от исходной *мощности* не имеет физических единиц), хотя единица измерения может быть просто «навязана» параметру-вершине природой участника; см. далее примеры с общепараметрическими именами *количество* и *коэффициент*.

Слабая метафоризация представляется одним из важных механизмов образования параметрических терминов — наряду с такими механизмами, как напр., опора на абстрактное общеколичественное слово, в принципе сохраняющее лексическое значение, если такое слово рассматривать изолированно внутри терминологического сочетания: *количество электричества*, *количество теплоты* (несмотря на специфичность сочетания в целом, все равно в обоих случаях это некое количество); аналогично *коэффициент теплопроводности* или *коэффициент трения* — все равно коэффициент, или как разнообразные характеризации физических явлений с помощью относительных прилагательных: *звуковое давление*, *динамическая вязкость* и др.

Для реализации задач обработки текста параметрические термины (устойчивые словосочетания, денотаты которых имеют индивидуальные свойства: *мощность множества* и др.) могут, в принципе, описываться отдельно, напр., в специальной терминологической зоне словаря или в онтологии; при этом их образование и функционирование как лексических единиц подчиняется общим семантическим закономерностям.

Приведенные примеры слабой метафоризации показывают, что ее степень, «дальность переноса» различна: *глубина паза* как величина, не утратившая буквальную пространственность, меньше отличается от исходной *глубины* (например, *глубины водоема*), чем *глубина просмотра сайта*, для которой пространственность уже метафорична, уже перешла в информационную сферу. Критерием отнесения обеих производных *глубин* в один класс, если ориентироваться на практические задачи типа Information extraction, может служить сохранение измеримости (так, *глубина просмотра сайта* тоже измеряется — например, количеством страниц, просмотренных одним пользователем за одно посещение).

2. Другим механизмом семантического переноса является метафоризация параметрического имени с утратой количественного измеряемого значения. В данной работе, где речь идет о количественных параметрах, будем называть такое явление «сильной метафоризацией», поскольку она выводит лексему за пределы собственно количественной сферы: *пульс планеты, пульс страны* — употребление не количественное. Механизм присущ лишь отдельным параметрическим именам; имя также должно обладать образностью, ассоциативным потенциалом (в данном примере действует ассоциация ‘пульс' — > 'жизнь /биологическая/’ — > ' жизнь социальная — события, деятельность, идеи, оценки и т. п.'). Среди параметрических существительных, для которых отмечается сильная метафоризация, имена *высота*, *глубина, объем, цена* и нек. др.: *высота поступка*, *глубина творческого замысла, глубина цвета, глубина изменений в структуре действия* (НКРЯ), *цена вопроса*, *объем работы* и т. п. Иногда утрачивается не только количественное параметрическое значение, но и параметрическое значение вообще; последнее диагностируется невозможностью употребления генитивной конструкции в контексте глаголов получения/передачи информации. Так, *объем исследований*, *объем хирургической операции*, на наш взгляд, утрачивает непосредственную измеримость, но не утрачивает параметричности (и абстрактной количественности тоже): такой *объем* можно *определить* или *указать* (быть может, он подлежит и неким косвенным измерениям: по количеству публикаций; путем перечисления затрагиваемых органов и тканей и т. п.), а, скажем, фразеологизм *глубина веков* полностью уходит из сферы параметров (**определить глубину веков*) и обозначает метафоризированную область пространства. При утрате параметричности генитивная конструкция может обозначать не параметр, а свойство или абстрактную сущность, с уже фиксированным большим значением параметра: *высота духа* (свойство, констатация; «высота» большая), *глубина души (/*в обороте *в глубине души/* — абстрактное вместилище; «глубина» также большая). В принципе, свойства с большим значением метафорически понимаемого параметра могут частично сохранять параметричность, которая выступает как абстрактная градуальность, например, *высота духа* у разных людей разная, и ее можно, например, *оценить* (но в отличие от полноценного параметра ее вряд ли можно *указать* в ответе на запрос). Таким образом, при метафорической утрате параметричности отмечается либо уход в (метафорические) пространственные ареалы, характеризующиеся большим значением одноименного параметра, либо констатация большого значения, быть может, частично градуируемого; семантическая нюансировка ’параметр — большое значение’ имеет место не только для исходных значений (см. Введение), но и для метафорических значений параметрических имен.

Альтернативность сохранения/утраты параметричности полисемичным словом вызывает ассоциацию с оппозицией 'параметрическая vs прямая диатеза' предикатов принятия решений: *выбрал жену* vs *выбрал в жены Марию* [Падучева 1999 а] (здесь фигурирует неколичественный параметр, имя социального отношения *жена*). Сопоставляемые явления относятся к несколько разным сферам семантики — в первом случае больше к лексической, а во втором,

описываемом в цитируемой работе — скорее, к семантике синтаксиса. В обоих случаях налицо некая двойственность, обусловленная особой информационной природой параметров, балансировкой между (параметрической) функцией и ее значением.

3. Еще одно явление, связанное с параметрическим именем — это метонимический переход в объекте-участнике от первичной для данного параметра таксономической категории к категории производной, «смежной», без изменения значения параметрического имени: *глубина погружения подводной лодки* = 'глубина того уровня, водного слоя, где находится погруженная лодка' (т. е. некоторой пространственной, предметной сущности*); возраст выхода на пенсию* = 'возраст людей, когда они обретают право выйти на пенсию'; *скорость поезда* = 'скорость движения поезда' (т. е. процесса), и т. п. У приведенных существительных очевидным образом просматриваются исходные категории участника: для *глубины* это вместилища, для *возраста* — живые существа, для *скорости* — движение. Их замена и происходит по механизму метонимии.

Данное явление широко распространено для параметрических имен; в ряде случаев, например для имен пространственных параметров, можно говорить о регулярных соответствиях таксономических классов, типа 'ПРЕДМЕТ (в широком смысле: вместилище, пространственный ареал, существо, геометрическая фигура и др.) — СИТУАЦИЯ (также в широком смысле: действие, движение, происшествие, природный процесс и др.)'; ср.: *высота дерева — высота полета*, *глубина колодца — глубина промерзания грунта, площадь комнаты — площадь возгорания*, *угол треугольника — угол вращения*. Фактически это одна из манифестаций фундаментальной онтологической оппозиции 'статика — динамика', отражающейся в системе языка.

Отметим, что метонимия здесь понимается расширительно, в соответствии с [Падучева 1999 б], когда изменения лексического значения предикатного слова может не быть, но происходит изменение так называемого ракурса видения обозначаемой ситуации (в нашем случае — ситуации характеризации объекта по параметру). Можно говорить о семантическом сжатии или об эллиптическом свертывании описания ситуации, о своего рода «метонимии-эллипсисе». Соответственно, смена ракурса при семантическом сжатии состоит в уменьшении «изображения», в отдалении «фотокамеры». Собственно, классическая метонимия (переименование по смежности, семиотическая «индексальность») тоже связана с процессом сжатия текста: сочетание *ложка меда* может пониматься как сокращение для фразы *наибольшее количество меда, которое вмещается в ложке*.

При рассмотрении производных контекстов (типа *скорость поезда*) правомерна задача экспликации исходной категории участника, задача «восстановления эллипсиса». Эта задача полезна, в частности, для повышения системности описания лексики в словаре. Например, исходная категория может приводиться первой или особо помечаться в перечне возможных. Для задачи извлечения параметрической информации указание в словаре исходной категории будет способствовать уточнению извлекаемых данных и обеспечивать возможность их интерпретирования и преобразований.

Может иметь место изменение значения участника. Например, имя действия в результате метонимического перехода может обрести предметное значение: в сочетании *длина пробега электрона* отглагольное имя *пробег* становится обозначением пространственного отрезка, который «пробегает» электрон. Похожую деривацию от S0 в сторону «предметной» пространственности претерпели, например, имя антропоцентричной единицы измерения *обхват,* параметрическое имя *размах* (*размах крыла самолета*), имя *поворот,* приобретшее способность обозначать фрагмент траектории.

Все же, как уже отмечено, классической метонимической смены значения у имени объекта обычно не происходит; так, в сочетании *глубина перекопки почвы* вряд ли имя *перекопка* приобретает предметное значение слоя почвы, до которого доходит лопата. Тут скорее чисто эллиптическая конструкция, незавершенный метонимический переход: предметное значение у имени *перекопка* как бы наметилось, но не развилось. Мы объединяем обе ситуации, когда изменение значения в участнике есть и когда оно не состоялось, и говорим о метонимии-эллипсисе. Подчеркнем ещё раз, что метонимия здесь, пусть и неполная, совершается не в параметрическом имени, а в имени участника; значение имени параметра неизменно.

Возможен, правда, и другой взгляд: в сочетаниях с отглагольным именем, типа *глубина перекопки*, первичным можно считать отглагольное имя, и тогда параметр выступает всего лишь участником глагольной ситуации: *перекапывать почву -> на какую глубину*. Но при таком взгляде сочетания *глубина борозды* или *лунки* (параметр — вершина) и *глубина перекопки* (параметр — «перевернутая» вершина, бывший «слуга» глагола) считались бы образованными по разным механизмам и были бы связаны между собой более опосредованно.

В профессиональной речи, в жаргоне могут иметь место завершенные «отраслевые» метонимические переходы, не свойственные обычному языку: *выплатить инфляцию клиенту* = выплатить сумму, возникшую в результате инфляции и не являющуюся настоящей прибылью; у слова *инфляция* (обычно обозначающего экономическое явление) здесь вырисовывается новое, окказиональное значение («денежная сумма»); встречается и метонимическое параметрическое употребление: *инфляция составляет 5,5%* (Интернет-материалы).

Имеются случаи, когда исходная категория не вполне очевидна. Например, для параметра *надежность* — его количественного варианта, обозначающего вероятность желаемого результата технологических процессов или действия устройств (соответственно, минус вероятность отказов и ошибок), представляется, что категория процессов (либо других их отглагольных «собратьев») первична по отношению к категориям предметным, ведь оцениваются не устройства как таковые, а их функционирование: *надежность электросетей* — ‘надежность работы электросетей’.

4. Среди семантических явлений, связанных с параметрическими именами, встречается и качественная оценка такого объекта — как правило, утилитарного, бытового, для которого точное количественное измерение (по крайней мере, по данному параметру) обычно не практикуется: *объем ресниц* (сочетание употребляется при рекламировании косметических средств), *глубина*

туфель, *сопротивление истиранию ламината* и подобное. Явление окказионально, встречается у нескольких «житейских» параметрических имен. Утрата измеримости связана не с метафоризацией, а, скорее, с десемантизацией, инволюцией значения измеряемого количественного параметра: качественная оценка -> количественный параметр — > качественная оценка.

5. Следует также отметить ситуацию 'ухода участника' [Падучева 1999 б], когда в определенном контексте (в профессиональной или в разговорной речи) участник очевиден, неспецифичен и потому может быть опущен. Такой эллипсис свойствен нескольким параметрам: *возраст*, *рост*, *пульс* (очевидно, что человека; например, в сочетании *переходный возраст* или в заголовке «*Зависит ли рост от питания?*» [Интернет-материалы]); *срок* (очевидно, что лишения свободы); *глубина* (крупного водоема), *частота* (радиосигнала) и некоторым другим. Конечно, при извлечении информации потребуется восстановление эллипсиса, для этого в словаре должна быть указана соответствующая прагматически значимая категория участника.

6. Среди семантических переходов следует назвать и образование по метонимии предметного значения у параметрических имен. Явление свойственно нескольким именам: *высота* (холм, «*…на безымянной высоте*»), *высота* (перпендикуляр в треугольнике), *сопротивление* (радиодеталь), *емкость* (резервуар), *площадь* (жилое помещение). Любопытно, что эти несколько примеров неоднородны. Так, у *высоты* (топографического объекта) и у *сопротивления* подразумевается повышенное значение одноименных параметров, а у *высоты треугольника*, *емкости* и *площади* такого компонента нет; последние образованы от параметров как функций, без фиксации количественного значения.

3. Вместо заключения. О путях дальнейших рассмотрений

Итак, указано несколько механизмов образования производных значений у имен количественных параметров. В данной работе при рассмотрении полисемии мы ограничились анализом наиболее представительной конструкции — генитивной. Однако приведенные семантические трансформации не исчерпывают картины полисемии. В схему переходов может быть включена и мотивирующая лексика, с ее семантическими преобразованиями. Например, сильно метафоризированная *высота* (*высота помыслов*) может описываться как дериват не от *высоты* — пространственного параметра, а от метафорического аксиологического значения прилагательного *высокий*. Такая схема была бы более точна, но (плата за точность) менее регулярна; повторное «вкрапление» прилагательного в семантическую деривацию параметрического существительного целесообразно только для отдельных отадъективных имен (*высота, глубина: высота стиля*, *глубина научного анализа* и нек. др.). Более полное описание полисемии, которое планируется получить в будущем, должно включить частеречные преобразования, с учетом индивидуальных особенностей лексем. Например, у прилагательного *высокий* прагматически важное производное общеколичественное значение не порождает номинативного коррелята; ср.: *высокая*

температура, но **высота температуры*; *высокая вероятность*, но **высота вероятности* (соответствующая реалия обозначается, например, именем *величина*). Известно и индивидуальное поведение пар *длинный — длина*, *широкий — ширина* в производных значениях: *длинный день*, но *долгота* (скорее, нежели *длина*) *дня*; *широкий кругозор*, но *широта* (а не *ширина*) *кругозора*. В подробной схеме деривации должно найтись место и параметрическим наречиям (*высоко*, *глубоко* и др.), с их переносными значениями (напр., порождающими такие формы, как *высокопоставленный*, *глубокоуважаемый*). Механизмы слабой и сильной метафоризации непосредственно от имени параметра, рассмотренные в данной работе «сквозь призму» генитивной конструкции, можно считать некоторым обобщением промежуточных частеречных «частностей», удобным для компактного словарного описания собственно параметрических имен и для практических задач типа Information extraction. Механизмы выпукло проявляются, например, на некоторых единицах неадъективной параметрической лексики: *объем памяти*, по-видимому, есть прямой дериват от физического *объема*. Хотя и для отдельных неадъективных имен возможны вкрапления переносных значений мотивирующей части речи: *вес политической фигуры* можно, в принципе, рассматривать как повторный морфологический дериват от *весить*, т. е. от *весить 2* («иметь влияние в обществе»). В целом картина полисемии при вовлечении неноминативной лексики получается довольно многообразная и требует отдельного, подробного картографирования.

Интересно также рассмотреть в дальнейшем сочетание, взаимодействие языковых и логических механизмов, механизмов математического и физического мышления. Например, термин *высота треугольника* есть результат следующих последовательных процессов:

1) параметризация психологической оценки формы треугольника, как бы лежащего горизонтально на одной из своих сторон: треугольник может восприниматься как «высокий» или «низкий»; эта вариативность осознается как параметр, значение которого требуется точно оценить;
2) формализация оценки — идея опустить перпендикуляр на горизонтально расположенное основание и измерить длину перпендикуляра;
3) слабая метафоризация (абстрагирование от строго вертикальной ориентации перпендикуляра);
4) образование предметного значения по метонимии; имя *высота* становится обозначением перпендикуляра.

В этом примере языковые механизмы метафоры и метонимии (за которыми стоят закономерности ассоциативного мышления) отчетливо проявляются на двух последних этапах; на первом же довлеет изначально вневербальное восприятие, а на втором функционирует геометрическое мышление.

В перспективе в рассмотрение также может быть включено сопоставление семантической деривации при формировании параметрической терминологии в русском и в других языках. Например, в английском языке математическое понятие *мощность множества* выражается конкурирующими терминами *potency, power, cardinal number, cardinality*, первые два из которых (как

и русский термин) прошли слабую метафоризацию от физического концепта «сила», а вторые два связаны с идеей обобщения и главенства, т. е. нефизической абстракции.

В целом такие рассмотрения должны способствовать изучению соотношений между обыденной и научной картиной мира.

Литература

1. *Апресян Ю. Д.* Лексикографические портреты (на примере глагола быть) // Научно-техническая информация. Сер. 2. — 1992, № 3.- С. 20–33.
2. *Курилович Е.* Очерки по лингвистике. Сб. статей. — М., 1962. — 456 с.
3. *Падучева Е. В.* Семантика глаголов выбора // ИАН СЛЯ. 1999. № 5–6. С. 34–42.
4. *Падучева Е. В.* О роли метонимии в концептуальных структурах // Труды Международного семинара Диалог'99 по компьютерной лингвистике и ее приложениям. В 2-х тт.- Т.1- Таруса, 1999. — С. 215–225.
5. *Падучева Е. В.* О параметрах лексического значения глагола: таксономический класс участника // Русский язык в научном освещении. 2002. № 1 (3). С. 87–111.
6. *Семенова С. Ю.* О таксономии актантов параметрических имен // Динамические модели : Слово. Предложение. Текст. — Сб. статей в честь Е. В. Падучевой. — М., «Языки славянской культуры», 2008. — С. 691–710.

ИССЛЕДОВАНИЕ ПЛАГИАТА В РАБОТАХ СТУДЕНТОВ

Шарапов Р. В. (info@vanta.ru),
Шарапова Е. В. (mivlgu@mail.ru)

Муромский институт (филиал) ФГБОУ ВПО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых», Муром, Россия

В работе проводится исследование студенческих работ на наличие заимствований (плагиата). Дается обзор подходов к сокрытию фактов заимствований. Проанализированы базы контрольных и курсовых работ, а также база студенческих статей.
Анализ показал, что для сокрытия фактов плагиата наиболее часто употребляются: изменение заимствованного текста, сокращение текста, корректировка родов, чисел и времен слов, замена русских букв на сходные по написанию латинские. При усилении контроля за оригинальностью работ увеличиваются попытки скрыть факт заимствований за счет изменения текста и увеличения числа используемых источников.

Ключевые слова: плагиат, обнаружение плагиата, заимствование, сокрытие плагиата, студент, замена букв, изменение текста, курсовая работа, контрольная работа

Введение

Бурное развитие вычислительной техники привели к глубокому проникновению компьютеров в нашу жизнь. Компьютеры окружают нас везде — на работе, дома, в магазинах и общественных местах. Современное развитие информационных технологий и глобальной сети Интернет предоставило широким кругам пользователей доступ к огромным массивам информации. Появилось большое число online-библиотек, содержащих в электронном виде художественную и научно-техническую литературу. Стало возможным читать книги, новости и газеты непосредственно с экрана компьютера.

В сети Интернет стало доступно множество методических указаний, курсов лекций, учебников и т. д. Кроме того, появились огромные коллекции рефератов, готовых лабораторных работ, курсовых и дипломных проектов и даже диссертаций. Использование компьютерной техники сильно облегчило задачу поиска и копирования подобной информации. Если раньше для написания реферата или контрольной работы информацию было нужно, по крайней мере, найти в книгах и переписать (вручную, перепечатать или ввести в компьютер

с помощью сканера и программ распознавания текстов), то теперь достаточно ввести название темы в поисковую систему и скопировать найденные материалы. Стал распространяться метод написания работ, получивший название «Copy & Paste». Метод заключается в простом копировании информации из одного или нескольких источников с минимальным редактированием получающегося таким образом текста.

Аналогичная ситуация наблюдается с отчетными материалами внутри учебных заведений. В связи с тем, что большое число пояснительных записок по курсовым и дипломным проектам выполняется с использованием компьютеров, происходит их распространение и повторное использование среди учащихся.

В последнее время наблюдается бурный рост использования в учебном процессе подобной заимствованной информации. Ситуация усугубляется тем, что учащиеся иногда не знают (не читают) то, что написано и «их» работах.

Плагиат — умышленное присвоение авторства чужого произведения науки или искусства, чужих идей или изобретений [1].

Как можно убедиться из определения, подобные заимствованные работы можно отнести к разряду плагиата. Задача обнаружения недобросовестного использования заимствованных текстов в учебных кругах (фактов плагиата) приобретает высокую актуальность. Не менее важной задачей следует считать всестороннее исследование подобных заимствований, что может помочь в их дальнейшем автоматизированном выявлении.

Постановка задачи

Среди всех видов заимствований, встречающихся в современном обществе, мы будем рассматривать только наиболее распространенный в студенческой среде — использование в своих работах чужих материалов без какого-то ни было собственного вклада. Чаще всего это простое копирование доступных материалов в свою работу, без их переработки (как это могло бы быть, например, при пользовании учебниками, курсами лекций или методическими указаниями). Иногда такой вид заимствований называют «неприкрытым копированием».

Цель работы — провести всестороннее исследование заимствований, встречающихся в студенческих работах (как учебных, так и научных). Интерес представляет то, как часто делаются попытки скрыть факт заимствований, какие подходы используются для этого и т. д.

Подходы к сокрытию заимствований

Существует большое количество подходов к сокрытию фактов плагиата. В связи с тем, что чаще всего работы проверяются на плагиат различными

компьютерными системами (чаще всего системой «Антиплагиат» [2]), подходы призваны обмануть эти системы. По мере развития и совершенствования систем проверки подходы также эволюционируют.

Для того, чтобы скрыть факт заимствований, студенты могут применяться следующие подходы [3]:

1. Корректировка родов, чисел и времён входящих в текст слов. Например, замена слова «выполнил» на «выполнила» или «выполнили», использование местоимения «я» вместо «мы» в оригинальном тексте и т. д.
2. Незначительное изменение заимствованного текста. Например, изменение по одному слову в предложении.
3. Сокращение заимствованного текста путем удаления слов, предложений, абзацев, рисунков, формул и т. д.
4. Перестановка частей текста, абзацев и предложений местами.
5. Обход систем проверки на плагиат путем замены русских букв на аналогичные по написанию английские буквы и т. д.
6. Замена знаков препинания: «.» на «,» и обратно, « » на «.» и т. д.
7. Замена пробелов на невидимые буквы (написанные, например, белым цветом).
8. Осуществление ручной или автоматической синонимизации текста.

Исходные данные

В качестве объекта исследований мы использовали три базы работ.

Первая база («Conference») включает в себя материалы, подаваемые на научную конференцию студентами факультета (студенты были заинтересованы в подаче работ). Конференция проходит раз в год. В базу были включены материалы за три последних года. Количество работ ежегодно варьировалось в пределах от 200 до 400. Объем каждой работы — до 1000 слов.

Таблица 1. Состав работ базы «Conference»

Год	Число поданных работ	Оригинальных работ	Плагиат
2009	402	182	220
2010	369	201	168
2011	205	108	97

В 2009 году студенты не знали о факте проверки работ системой «Антиплагиат». В 2010 и 2011 году они были предупреждены заранее. Поэтому количество оригинальных работ увеличилось. С другой стороны, в работах стали применяться различные подходы к сокрытию фактов плагиата. Надо заметить, что под оригинальными работами мы понимаем работы, для которых не было найдено соответствия в сети Интернет и распространенных учебниках. Для оценки оригинальности кроме системы «Антиплагиат» использовались

и другие методы, в том числе авторская разработка «Автор.NET» [4]. Тем не менее, нельзя исключать тот факт, что некоторые работы, признанные оригинальными, все же могут содержать скрытые заимствования.

Вторая база работ («Course») включает в себя базу курсовых работ, выполненных студентами кафедры. Объёмы работ варьируются от 25 до 100 страниц.

Третья база («Control») состоит из контрольных работ, выполненных студентами заочной формы обучения по одной из гуманитарных дисциплин. Объёмы работ варьируются от 5 до 50 страниц.

Исследование подходов к сокрытию заимствований

Анализ базы «Conference» показал, что доля работ, применяющих различные подходы к сокрытию фактов заимствований, ежегодно растет (см. таблицу 2). Это объясняется, видимо, начавшейся борьбой с заимствованиями в студенческих работах.

Таблица 2. Процент работ в базе «Conference», скрывающих факты заимствований

Год	Работы, содержащие заимствования	Работы, скрывающие факт заимствований	Процент работ, скрывающих факт заимствований
2009	220	14	6,4%
2010	168	135	80,4%
2011	97	91	93,8%

Анализ показал, что подавляющее большинство работ копируются из одного источника (см. таблицу 2), чаще всего из учебников, статей из сети Интернет и публикаций региональной прессы. Тем не менее, число работ, компонуемых из двух и долее источников, ежегодно возрастает (см. таблицу 3). В 2011 году их доля достигла 15%. Это связано, вероятно, со знанием авторов о том, что их работы будут проверяться на плагиат.

Таблица 3. Количество источников для заимствований в работах базы «Conference»

Количество источников	Доля в 2009 г., %	Доля в 2010 г., %	Доля в 2011 г., %
Копирование текста из одного источника	98%	91%	85%
Копирование и компоновка текста из нескольких источников	2%	9%	15%

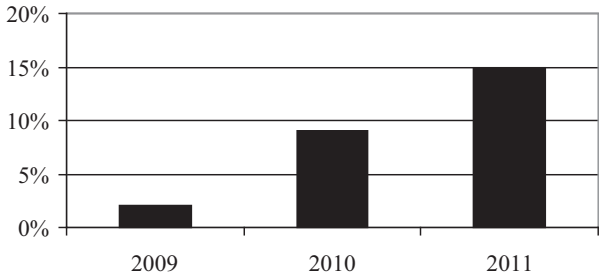


Рис. 1. Доля работ базы «Conference», скопированных и скомпонованных из нескольких источников

В связи с тем, что наиболее распространенной и доступной (как преподавателям, так и студентам) системой обнаружения заимствований является «Антиплагиат», большинство попыток скрыть факт заимствований нацелено именно на эту систему.

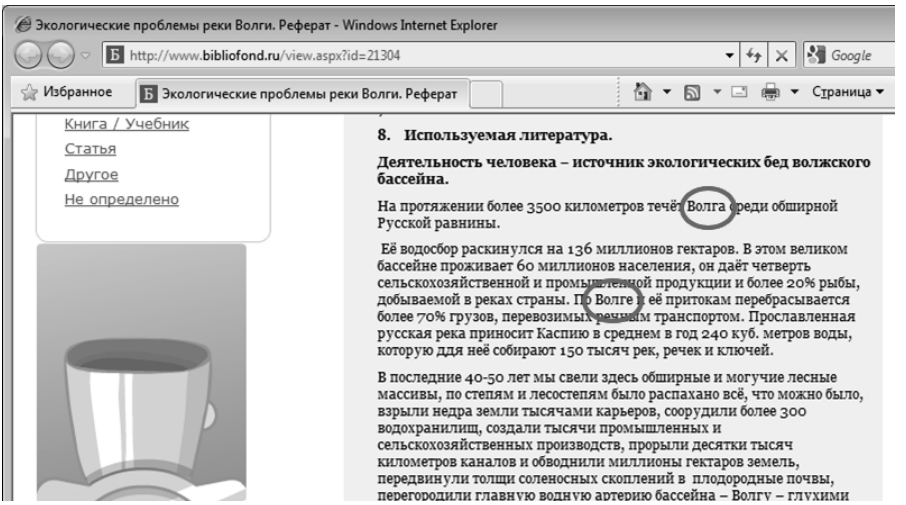
Анализ подходов, используемых студентами для сокрытия факта плагиата, показал, что в 33 % работ осуществлялась корректировка родов, чисел и времён слов (см. таблицу 4). Наиболее применяемый студентами подход к сокрытию факта плагиата — это незначительное изменение текста. Его доля достигла 89 % в работах 2011 года. Так, например, делались вставки слов и предложений в заимствованный текст, подвергались изменению названия населенных пунктов и рек (р. Волга в оригинале заменялась на р. Ока в статье). Надо заметить, что часть работ кроме заимствованных текстов содержала оригинальные блоки, чаще всего введение и заключение. Приведенная выше доля статей, подвергавшихся изменению, учитывает только заимствованные части таких текстов.

Экологические проблемы Оки

На протяжении более 3500 километров течёт Ока среди обширной Русской равнины.

Её водосбор раскинулся на 136 миллионов гектаров. В этом великом бассейне проживает 60 миллионов населения, он даёт четверть сельскохозяйственной и промышленной продукции и более 20% рыбы, добываемой в реках страны. По Оке и её притокам перебрасывается более 70% грузов, перевозимых речным транспортом. Прославленная русская

а)



б)

Рис. 2. Пример незначительного изменения текста
а) В проверяемой работе указана р. Ока
б) В оригинале указана р. Волга

Таблица 4. Частота использования подходов к сокрытию фактов плагиата в работах базы «Conference»

Подходы к сокрытию плагиата	Доля в 2009 г., %	Доля в 2010 г., %	Доля в 2011 г., %
Корректировка родов, чисел и времён слов	28,6%	33,3%	33%
Незначительное изменение текста	72%	70,3%	89%
Сокращение заимствованного текста	28,6%	43,7%	37,3%
Перестановка частей текста	0%	3%	2,2%
Замена букв	7%	25,9%	6,6%
Замена знаков препинания	0%	1,5%	1,1%
Замена пробелов на невидимые буквы	0%	0,7%	0%
Синонимизация текста	0%	1,5%	1,1%

Из работ, скопированных из одного источника, около 40% подвергались сокращению. В данном случае под сокращением подразумевалось исключение части предложений, графиков, рисунков из заимствованных текстов, а также исключение начальных или конечных блоков текста, по смыслу составляющих единое целое с заимствованным фрагментом. Копирование законченного фрагмента из текста (например, раздела или главы) сокращением не считалось.

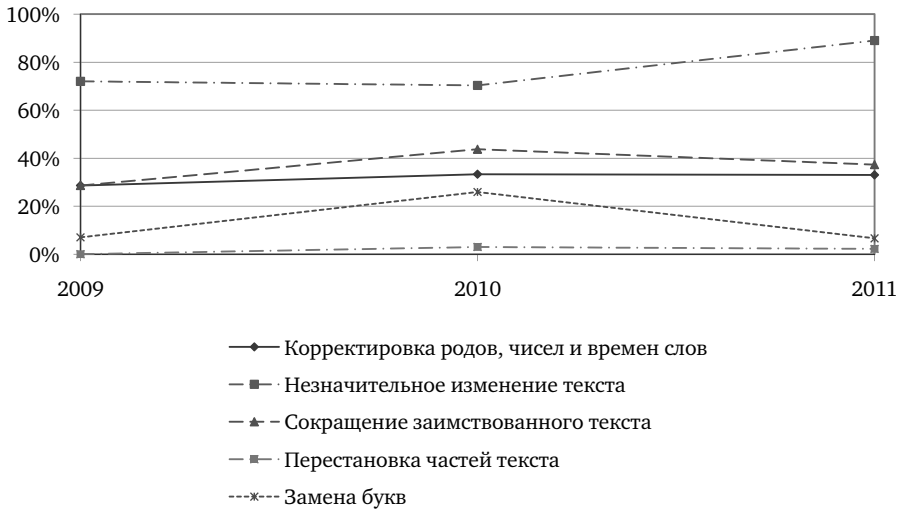


Рис. 3. Доля наиболее используемых подходов к сокрытию фактов плагиата в базе «Conference»

Замена букв активно использовалась в работах 2010 года и была обнаружена при проверке. Вследствие этого, в работах 2011 года доля замены букв снизилась до 6,6%. Чаще всего замене подверглись русские буквы «о», «е», «а», сходные по написанию с английскими буквами.

Сегодня воздействие человека на окружающую среду приняло угрожающие масштабы. Поэтому охрана природы – задача нашего века, проблема, ставшая социальной. Наше здоровье, благополучие и развитие человеческой цивилизации в целом зависят только от нас и наших действий Мы считаем, что достижение стабильной экологической обстановки станет возможным тогда, когда люди осознают, «что все взаимосвязано со всем», и природа всегда будет давать нам то, что мы сумели дать ей.

Рис. 4. Результаты проверки орфографии средствами Microsoft Word текста, содержащего замены русских букв «о», «е», «а» на сходные по написанию английские буквы

Замена знаков препинания («.» на «,» и обратно) осуществлялась в 1% работ. Замена пробелов на невидимые буквы написанные белым цветом обнаружена всего в одной работе. Ручная синонимизация проводилась только в 1% работ. Применения автоматической синонимизации в статьях замечено не было. В целом, последние три подхода применяются студентами достаточно редко.

Объем работ в базах «Course» и «Control» существенно больше, чем в базе «Conference». По этой причине обнаружение фактов заимствований в них — задача более сложная и трудоёмкая. Например, система «Антиплагиат» ограничивает объём проверяемого текста 3000 или 5000 символами (доступно после регистрации). В наше рассмотрение попали работы со значительной долей заимствований.

Следует отметить, что при сдаче указанных работ, проверка преподавателями оригинальности не проводилось. Этот факт оказал значительное влияние на содержание работ.

Анализ показал, что среди курсовых работ (база «Course) доля заимствований из сети Интернет достаточно мала — в основном это введение и теоретическая часть. С другой стороны, достаточно сильно работы пересекаются между собой. Другими словами заимствование происходит у своих же одногруппников или из работ старших курсов. Доля работ, скопированных из одного источника, составляет 78% (см. таблицу 5). Для того, чтобы скрыть факт заимствований, в основном применяется изменение визуального представления работы (оформления) — изменение шрифтов, интервалов, разбиение на абзацы и т.д. Иногда части работ переставляются местами. В большинстве работ (89%) производилась незначительная корректировка текстов (см. таблицу 6). Это объясняется в первую очередь «подгонкой» текстов под задание на курсовую работу.

Таблица 5. Количество источников для заимствований в работах баз «Course» и «Control»

Количество источников	**Курсовые работы, %**	**Контрольные работы, %**
Копирование текста из одного источника	78%	58%
Копирование и компоновка текста из нескольких источников	22%	42%

Таблица 6. Частота использования подходов к сокрытию фактов плагиата в базах работ «Course» и «Control»

Подходы к сокрытию плагиата	**Курсовые работы, %**	**Контрольные работы, %**
Корректировка родов, чисел и времен слов	14%	8%
Незначительное изменение текста	89%	76%
Сокращение заимствованного текста	17%	32%
Перестановка частей текста	3%	0,5%
Замена букв	0%	0%
Замена знаков препинания	0%	0%
Замена пробелов на невидимые буквы	0%	0%
Синонимизация текста	0%	0%

В базе контрольных работ «Control» ситуация несколько иная. 42% работ, замеченных в заимствованиях, копировались и компоновались из нескольких источников. Наиболее часто употребляется незначительное изменение текста (76% случаев), сокращение текстов (32%) и корректировка родов, чисел и времен слов (8%).

Выводы

Исследования показали, что для сокрытия фактов плагиата наиболее часто употребляются:

- изменение заимствованного текста,
- сокращение текста,
- корректировка родов, чисел и времен слов,
- замена русских букв на сходные по написанию латинские.

Такие подходы, как перестановка частей текста, замена знаков препинания, замена пробелов на невидимые буквы, ручная и автоматическая синонимизация текста применяются достаточно редко.

При усилении контроля за оригинальностью работ увеличиваются попытки скрыть факт заимствований за счет изменения текстов и увеличения числа используемых источников.

Подходы, используемые для маскировки плагиата, со временем могут меняться в зависимости от их эффективности против систем проверки на плагиат.

Литература

1. *Плагиат* — Википедия [Электронный ресурс]. — Режим доступа: http://ru.wikipedia.org/wiki/Плагиат
2. *Антиплагиат* [Электронный ресурс]. — Режим доступа: http://www.antiplagiat.ru/
3. *Шарапов Р. В.* Анализ подходов к обнаружению заимствованных текстов // Журнал «Современные наукоемкие технологии» — М: Российская академия естествознания, 2011 г. № 3, С. 47–49
4. *Шарапов Р. В., Шарапова Е. В.* Система проверки текстов на заимствования из других источников // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: XIII Всероссийская научная конференция «RCDL'2011». Воронеж, 19–22 октября 2011 г.: труды конференции — Воронеж: Издательско-полиграфический центр Воронежского государственного университета, 2011, с. 233–238

КОРПУС ИЛИ ЭКСПЕРИМЕНТ?[1]

Шмелев А. Д. (shmelev.alexei@gmail.com)

Московский педагогический государственный университет, Москва, Россия; Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

В статье обсуждается общенаучное разграничение наблюдения и эксперимента применительно к лингвистическим исследованиям.
Утверждается, что некоторые лингвистические задачи должны решаться на основе наблюдения, т. е. посредством обращения к корпусным данным, тогда как для решения других необходим лингвистический эксперимент; при этом исследователь может проводить эксперимент как с опорой на собственную языковую компетенцию, так и путем опроса информантов.
Описаны проблемы, связанные с проведением опроса информантов для выявления региональных языковых норм на фоне общеязыковых норм с одной стороны и регионального узуса - с другой.

Ключевые слова: наблюдение, эксперимент, корпус, узус, языковая норма, региональная норма

1. Наблюдение и эксперимент: общенаучное разграничение

Дискуссия на конференции «Диалог-2011», посвященная использованию корпусов в лингвистических исследованиях, показала, что ряд участников дискуссии был готов характеризовать исследования, проведенные на основе «Национального корпуса русского языка», как разновидность экспериментальных исследований. Представляется, что такое словоупотребление находится в противоречии с общенаучным разграничением двух основных способов получения данных: наблюдения и эксперимента. Различие между ними может быть описано следующим образом. Наблюдение имеет своим объектом явления и события, происходящие независимо от исследователя, который тем самым изучает их in vivo. Напротив того, эксперимент предполагает целенаправленную деятельность, направленную на получение данных, которые без эксперимента просто не существовали бы; тем самым производится исследование in vitro. Существенно при этом иметь в виду ограниченность экспериментального метода. Прежде всего, существуют вещи, которые в принципе нельзя исследовать экспериментально, так что для их изучения приходится ограничиваться

[1] Статья написана при финансовой поддержке РГНФ в рамках научно-исследовательского проекта РГНФ («Кодифицированная норма и региональный узус»), проект № 11-04-00501а.

наблюдением[2]. Так, климатические условия в той или иной местности, по-видимому, не подлежат экспериментальному изучению. Едва ли имеет научный смысл экспериментальная проверка таких гипотез, как «если я забываю зонтик, идет дождь, а если я беру его с собою, светит солнце», а более содержательные эксперименты с климатом едва ли возможны (имеются в виду научные эксперименты, а не воздействие на климат как побочный эффект человеческой деятельности). Точно так же экспериментальные методы в собственном смысле слова не могут использоваться в астрономии. Кроме того, чрезвычайно важно, что если часто приходится ограничиваться наблюдением, не прибегая к экспериментам, то эксперимент без наблюдения вообще невозможен. Во-первых, эксперименту, как правило, предшествует наблюдение, на основании которого и выдвигаются экспериментально проверяемые гипотезы. Во-вторых, что важнее, наблюдение над полученными экспериментальными данными составляет неотъемлемую часть всякого эксперимента; в противном случае проведение эксперимента лишилось бы смысла. Если же оказывается, что исследователь не нуждается в экспериментальных данных (либо проведение эксперимента невозможно в принципе) и ограничивается данными, полученными в ходе наблюдений, он прибегает к их обработке, но такая обработка не должна называться «экспериментом».

2. Наблюдение и эксперимент: применение общенаучного разграничения к лингвистическим исследованиям

Все сказанное имеет непосредственное отношение и к лингвистическим исследованиям. Многие виды лингвистических исследований заведомо не предполагают и не могут предполагать эксперимента. Так, при изучении древних языков приходится ограничиваться наблюдением над данными, извлекаемыми из корпуса имеющихся в распоряжении исследователя текстов на этих языках. Разумеется, в какой-то момент могут быть обнаружены новые тексты, и они дадут материал для проверки описания, созданного на основе предшествующего корпуса; однако это не будет экспериментальной проверкой, поскольку имеет место именно обнаружение, а не искусственное создание новых данных. Точно так же и астрономические наблюдения могут дать материал для проверки существующих теорий, однако это не делает астрономические наблюдения экспериментом[3].

Изучение живой устной речи также в норме должно вестись in vivo, т. е. посредством наблюдения, а не эксперимента. Возможно, впрочем,

---

[2] Для простоты я отвлекаюсь от возможности косвенных экспериментов с объектами, моделирующими изучаемые условия. В любом случае таким образом нельзя получить непосредственные данные, касающиеся интересующего нас объекта.

[3] Само собою разумеется, что не является экспериментом ситуация, когда исследователь конструирует некоторое языковое выражение, а потом проверяет, обнаружится ли оно в корпусе древних текстов.

экспериментальное исследование спонтанного речевого поведения носителей языка в некоторой искусственно созданной ситуации.

Однако следует понимать, что есть языковые данные, которые можно получить только экспериментально, поскольку непосредственному наблюдению они недоступны. (Из сказанного вытекает, что для древних языков эти данные нельзя получить в принципе, хотя иногда их можно реконструировать на основе косвенных соображений.) Так, важнейшим элементом языковой способности носителей языка является их умение отличить «правильное» высказывание от «неправильного». Необходимо добавить, что во многих случаях бинарное противопоставление «правильное»– «неправильное» оказывается недостаточным, так что приходится вводить какие-то промежуточные точки шкалы. При этом следует иметь в виду, что «промежуточный» статус на шкале «правильности» того или иного высказывания может быть обусловлен двумя различными обстоятельствами. Во-первых, сами информанты могут колебаться в оценке степени правильности высказывания, характеризовать его как «допустимое, но странное», «сомнительное» и т. п. Во-вторых, иногда обнаруживается разнобой во мнениях информантов: одни считают высказывание правильным, другие — нет. В обоих случаях одно наблюдение над корпусом не позволит выявить такие высказывания. В самом деле, может случиться так, что информанты, считающие, что высказывание правильное, будут употреблять его, тогда как мнение других, возможно, даже более многочисленных информантов, отвергающих высказывание как неправильное, выявлено не будет. Сомнительные высказывания также могут встретиться в корпусе, и их статус не будет отличаться от прочих, несомненно правильных, высказываний.

Здесь мы сталкиваемся с более общей проблемой, касающейся наблюдения над готовыми текстами. Исследователи, имеющие дело с древними текстами, знают, что в них могут встречаться ошибки разного рода, но исходят из того, что ошибка не может быть частотной. Однако, имея дело с современными языками, мы нередко обнаруживаем, что носители языка пользуются понятием «распространенная ошибка». Понятно, что одно наблюдение над текстами не позволит отличить такую ошибку от «правильных» (т. е. не вызывающих ни малейшего сомнения у носителей языка) высказываний.

В силу сказанного для древних языков оказывается затрудненным разграничение нормы и узуса. Возможны лишь разного рода ухищрения, позволяющие в каких-то случаях проводить такое разграничение на основе косвенных данных (так, Клайв Льюис отмечал, что если в старых текстах делается предостережение против некоторой ошибки, то это значит, что соответствующее речевое явление было довольно распространено в узусе).

В отношении современных языков многие из указанных трудностей могут быть преодолены при помощи эксперимента. Самый примитивный эксперимент состоит в опросе информантов, которым предъявляется высказывание и делается предложение оценить это высказывание с точки зрения языковой правильности («можно ли так сказать?»). Однако следует иметь в виду, что результаты такого эксперимента вовсе не обязательно будут отражать лингвистическую реальность (т. е. в данном случае подлинное мнение информантов

о правильности предъявленных высказываний). Подводные камни, связанные с проведением опросов, известны специалистам, работа которых во многом опирается на опрос как метод исследования (не только полевым лингвистам, но и, напр., социологам), и я не буду здесь на них останавливаться.

Возможны и другие типы опросов. Так, можно моделировать какую-то ситуацию и наблюдать за речевым поведением испытуемых в данной ситуации. При этом прямолинейный способ получения ответа состоит в том, чтобы эксплицировать параметры ситуации, предъявить их респонденту, и спросить, что он сказал бы в такой ситуации. Более изощренный способ заключается в том, чтобы искусственно поставить испытуемого в некоторые условия и наблюдать за его речевым поведением,– это приближает искусственно созданные условия к естественным и тем самым эксперимент оказывается похож на наблюдение in vivo, но все же отличается от него.

Здесь, впрочем, важны формулировки вопросов. Если мы спрашиваем респондента, что бы он сказал в некоторой ситуации, то предполагаем получить данные о том, как представляет себе респондент собственное речевое поведение (излишне говорить, что не всегда эти представления оказываются адекватными). Можно спросить, что респондент считает правильным сказать в данной ситуации или что обычно говорят в такой ситуации, и получить другие ответы.

Помимо этого, в применении к живым языкам возможно непосредственное наблюдение над повседневной речевой деятельностью носителей языка, а не только наблюдение над результатами такой деятельности — готовыми текстами. Известны трудности, связанные с таким наблюдением: поведение людей, ставших объектом наблюдения, может меняться в силу того, что им становится известно, что за ними наблюдают[4], а также разные приемы, направленные на преодоление этих трудностей (так, для изучения речи некоторых социальных групп может использоваться так называемое «включенное наблюдение» и т. п.).

Заметим, что то, что говорилось о соотношении наблюдения и эксперимента вообще, в целом применимо и к лингвистике. В частности, и в лингвистике, прежде чем проводить опрос информантов, конструировать искусственные примеры, придумывать воображаемые ситуации, исследователь обычно обращает внимание на какие-то явления, с которыми он столкнулся, наблюдая повседневную речевую деятельность, а любой эксперимент включает в себя наблюдение в качестве составной части. Наконец, можно заметить, что если в результате эксперимента был получен значительный массив данных, эти данные могут составить некий корпус, над которым другие исследователи могут проводить наблюдения, уже не занимаясь экспериментами.

[4] Подобные трудности иногда возникают и в естественных науках: так, при наблюдении за поведением животных в естественных условиях важно «не спугнуть» животное.

3. Роль интроспекции

Одним из методов лингвистического эксперимента является интроспекция. Внутренняя форма этого слова как будто подсказывает, что интроспекция — это разновидность наблюдения (наблюдение над самим собою). Однако интроспекция как чистое наблюдение встречается довольно редко. Конечно, бывает так, что, сказав или написав нечто без целей лингвистически исследовать соответствующее языковое явление, лингвист про себя отмечает, что именно он сказал или написал и в дальнейшем использует это наблюдение в своей работе. Однако гораздо чаще интроспекция включает эксперимент: лингвист искусственно конструирует некоторое высказывание (часто переставляя или заменяя какие-то языковые единицы, которые он уже наблюдал в речевой деятельности) и оценивает его правильность. Другой тип эксперимента в рамках интроспекции — попытка лингвиста представить себе, что бы он сказал в некоторой воображаемой ситуации.

Интроспекция обладает целым рядом важных преимуществ по сравнению с работой с информантами. Однако необходимо понимать, что, ограничиваясь интроспекцией, лингвист замыкается на собственном идиолекте, который может по каким-то характеристикам отличаться от идиолектов других носителей языка. Помимо этого, эксперимент, сводящийся к интроспекции, в большей степени ориентирован на норму, а не на узус; конечно, относительно каких-то высказываний, которые представляются ему неправильными, лингвист может понимать, что неоднократно слышал или читал (а возможно, и сам произносил) их, однако в точности оценить степень их распространенности в узусе путем интроспекции невозможно. Поэтому, если целью исследования является изучение узуса (напр., количественное), интроспекция мало что дает.

Все сказанное касается и проблемы соотношения нормы и узуса, а также более специальных вопросов, напр. соотношения региональной нормы и регионального узуса с общеязыковой нормой и общеязыковым узусом. Здесь исследователь, даже живущий в соответствующем регионе, не может ставить во главу угла интроспекцию. Региональный узус (и его соотношение с общеязыковым) в принципе может изучаться посредством наблюдения (хотя, напр., для изучения произносительного узуса нужен был бы представительный корпус устной речи), но региональная норма в ее соотношении с общеязыковой нормой требует обращения к работе с информантами. В этом отношении первостепенную роль играет корректная методика опросов и интерпретация результатов.

4. Отклонение регионального узуса от общеязыковой нормы: проблемы интерпретации и способы исследования

При изучении соотношения общеязыковой нормы и регионального узуса в фокус внимания исследователя должны попасть случаи, когда региональный узус отличается от того, что предписывается кодифицированной общеязыковой

нормой. Для простоты будем исходить из того, что словари и справочники, кодифицирующие общеязыковую норму, во всех случаях содержат адекватное описание (хотя на самом деле это, конечно, не так). Тогда отличия регионального узуса от того, что предписывается нормативными словарями и справочниками, могут характеризоваться как отклонения регионального узуса от общеязыковой нормы.

Такие отклонения могут интерпретироваться различным образом в зависимости от того, каковы в данном случае общеязыковой узус и региональная норма. Логически мыслимы четыре ситуации:

1. Как общеязыковой узус, так и региональная норма соответствует общеязыковой норме. В этом случае мы можем говорить о специфических явлениях в региональном узусе, которые не становятся нормативными даже в соответствующем регионе.
2. Общеязыковой узус соответствует общеязыковой норме, а региональная норма допускает употребление, характерное для регионального узуса. В этом случае можно говорить о специфической региональной норме, отличной от общеязыковой нормы.
3. Региональная норма соответствует общеязыковой норме, а в общеязыковом узусе обнаруживаются то же отклонение, что и в региональном узусе. В этом случае можно говорить об отсутствии региональной специфики: соответствующее отклонение от нормы обнаруживается не только в рассматриваемом регионе, но и распространено более или менее повсеместно.
4. Отклонение от общеязыковой нормы обнаруживаются не только в региональном, но и в общеязыковом узусе; при этом данное отклонение допускается региональной нормой. В этом случае некоторое распространенное повсеместно отклонение от (общеязыковой) нормы «повышает» свой статус в региональной разновидности языка.

Замечание. Понятно, что языковое выражение, употребительное или даже известное лишь в некотором регионе, является отклонением от общеязыковой нормы лишь при наличии альтернативного и при этом общеязыкового способа выражения. Очевидный пример — региональная топонимика. Вероятно, наименование *Столбы* известно не только жителям Красноярска; но даже если бы его в других регионах не знали, оно все равно оставалось бы нормативным с точки зрения общеязыковой нормы (разумеется, это верно не для всех региональных топонимов). Особо следует упомянуть случаи, когда та или иная реалия получает разные наименования в различных регионах либо, наряду с общеязыковым наименованием, у него имеется особое «региональное» наименование. Часто бывает так, что региональное наименование является общеязыковой нормой, когда используется для обозначения данной реалии в соответствующем регионе (напр., для придания местного колорита), тем более и реалии в разных регионах часто не тождественны. Так, сельский дом на севере России именуется *изба*, а на юге — *хата*. Соответственно, северные бревенчатые дома уместно называть *избами*, а южные дома (как правило, выполненные

не из цельных, а из половинных бревен, снаружи обмазанные глиной и побеленные) — *хатами*. Слова *арык* или *ишак* (вместо слов *канал* и *осел*) уместны, когда речь идет о Средней Азии; так же, если речь идет о Средней Азии, вместо «общеязыкового» слова *абрикос* (в применении к дереву) может использоваться *урюк* (при этом, поскольку носители русского языка часто именно в Средней Азии сталкиваются с цветущими абрикосовыми деревьями, сочетание *цветущий урюк* оказывается едва ли не более частотным, нежели *цветущий абрикос*)[5].

Однако, для того чтобы отнести конкретную ситуацию к тому или иному типу необходимо уметь устанавливать в каждом конкретном случае общеязыковую и региональную норму и общеязыковой и региональный узус. Как уже говорилось, общеязыковую норму можно условно приравнять к рекомендациям нормативных словарей и справочников. Как общеязыковой, так и региональный узус, вообще говоря, следовало бы изучать методом наблюдения. Поскольку речь идет о современной языковой ситуации, возможно не только наблюдение над готовыми результатами речевой деятельности (корпусом текстов), но и непосредственное наблюдение над речевой деятельностью в прямом режиме (последнее — если исследуется устный узус). Именно наблюдение над речевой деятельностью в прямом режиме часто является исходным стимулом для последующих наблюдений над корпусом текстов и лингвистическими экспериментами. Однако, как уже говорилось, непосредственное наблюдение в прямом режиме сопряжено с риском, что сам факт наблюдения будет воздействовать на речевое поведение носителей языка; кроме того, оно плохо применимо к созданию письменных текстов. В любом случае результаты непосредственного наблюдения в прямом режиме затруднительно использовать для решения многих важных исследовательских задач — в частности, сопряженных со статистическим анализом полученных результатов (так, оценка частотности того или иного произносительного варианта при непосредственном наблюдении может делаться лишь «на глазок»). По отношению к письменным текстам часто удобным оказывается «сегментно-статистический анализ» [Беликов 2011], использующий в качестве источника корпуса текстов различные сегменты русского Интернета. Однако для исследования особенностей регионального узуса, специфичных для устной речи, данные Интернета мало что дают. Попытка получить материал для наблюдения методами полевой лингвистики (в духе проекта петербургских лингвистов «Один речевой день» [Богданова 2008]) во многих случаях также не дает и едва ли может дать в обозримом будущем достаточный материал для статистически достоверных выводов, касающихся особенностей общеязыкового и регионального узуса. Здесь более эффективным может оказаться «скрытый» эксперимент, когда искусственно создается ситуация, в которой велика вероятность употребления информантом языкового выражения, интересующего исследователя. Что касается до региональной нормы, то в отношении письменного языка она может изучаться методом наблюдения — если принять положение, согласно которому определенные типы текстов заведомо написаны нормативным языком (напр., официальные

[5] Ср. обсуждение этого вопроса в статье [Shmelev, Shmeleva 2009, 340].

постановления местных органов власти). Однако допущения такого рода в отношении устной нормы сильно исказило бы картину. В самом деле, такая методика предполагала бы, что мы a priori выделяем некоторое заданное множество «образованных» носителей языка данного региона, речь которых может считаться «образцовой», и постулируем, что любое языковое явление, которое встретилось в их речи, является принадлежащим норме (хотя бы региональной). Однако такой подход был бы неоправданным упрощением: норма была бы приравнена к узусу «образованных» носителей языка, при этом случаи, когда некоторое языковое явление, встретившееся в речи одних «образованных» носителей языка, решительно отвергается другими, вообще не нашли бы отражения в описании.

5. Изучение соотношения нормы и узуса путем опроса информантов: подводные камни

Сказанное подводит нас к выводу, что здесь, как и в других подобных случаях, не обойтись без эксперимента. При этом, как уже говорилось, чрезвычайно важна корректная методика эксперимента и интерпретации его результатов. Иллюстрируем это на примере произносительных норм, отклонения от которых наблюдаются в узусе.

Предположим, исследователь в ходе наблюдений обнаружил, что некоторая словоформа в региональном узусе часто произносится не с тем ударением, которое предписывается общеязыковой нормой (определяемой, напомним, на основе нормативных словарей и справочников). Правильная интерпретация этого факта зависит от того, насколько частотно такое отклоняющееся от нормы ударение в региональном узусе, встречается ли оно в региональном узусе чаще, нежели в общеязыковом узусе, осознается ли оно как отклонение от нормы «образованными» носителями языка, живущими в данном регионе.

Если для данного региона исследователь располагает представительным корпусом устной речи, в котором интересующая его словоформа встречается достаточно часто, чтобы можно было сделать какие-то статистически значимые наблюдения, то дополнительных экспериментов для изучения регионального узуса не потребуется. В противном случае, чтобы получить информацию о том, как в данном регионе произносят интересующую нас словоформу, требуется провести эксперимент. Методологически предпочтителен «скрытый» эксперимент, когда создается ситуация, которая естественным образом будет располагать информанта к произнесению данной словоформы. Однако иногда оказывается, что экономнее просто дать информантам прочесть вслух текст, уже содержащий данную словоформу. Разумеется, в последнем случае произношение информантов может быть несколько иным, нежели в естественных условиях, в большей мере ориентированным на их представления о норме, но этим часто можно пренебречь. К более существенным искажениям может привести то, что при проведении такого эксперимента способ произнесения каждого

информанта считается один раз, независимо от того, использует ли он соответствующее слово в своей повседневной речи. Некоторая поправка могла бы состоять в том, чтобы не считать ответы информантов, очевидным образом испытывающих затруднение при произнесении данной словоформы (об этом может свидетельствовать заминка при чтении). Впрочем, на практике при проведении такого эксперимента все равно редко достигается репрезентативность отбора информантов, так что возможными сдвигами можно пренебречь.

Для установления региональной нормы в рассматриваемой области трудно представить себе более эффективный способ, нежели опрос информантов. При установлении орфоэпической нормы следует ориентироваться не на то, как носители языка произносят единицу, для которой необходимо выработать орфоэпическую рекомендацию, а на их сознательную оценку степени нормативности существующих орфоэпических вариантов [Шмелев 2004], и это касается как общеязыковой, так и региональной нормы. Иными словами, тот факт, что какое-то ударение, отвергаемое общеязыковой нормой, широко распространено в данном регионе, в том числе в речи радиодикторов или университетских преподавателей, еще не означает, что его следует счесть региональной нормой: другие дикторы и преподаватели (или даже большинство таковых) могут это ударение решительно отвергать. Решение вопроса, имеем ли мы дело с региональным узусом, отклоняющимся от нормы, или с особой региональной нормой, не может быть принято только на основе фактов узуса: необходимо учитывать оценку этих фактов говорящими. Так, в Америке почти все русские говорят (или, по крайней мере говорили несколько лет тому назад) *на Интернете*, а не *в Интернете*; необходимо специальное исследование, чтобы решить, как трактовать это: как ошибку, вызванную отрывом от родной речи, или как особую черту русского языка в Америке.

При проведении опроса важно найти оптимальную формулировку вопросов и правильно интерпретировать полученные ответы. Так, может ввести в заблуждение формулировка вопроса: «Какое ударение представляется вам наиболее привычным?» Респонденту может быть не вполне ясно, спрашивают ли его о том, какое ударение он привык слышать, или о том, как он сам привык произносить данное слово. Целесообразно разграничить три типа вопросов:

(1) С каким ударением обычно произносят это слово окружающие вас люди?
(2) С каким ударением произносите это слово вы?
(3) Как правильно произносить данное слово?

Желательно каждому респонденту дать все три вопроса, сделав тем самым различие между ними максимально явным. Помимо этого полезно предусмотреть возможность «уклончивых» ответов: на первый вопрос — «не обращал внимания» или «произносят по-разному», на второй вопрос — «не использую этого слова», на третий вопрос — «не знаю, надо посмотреть в словаре». Наличие возможности дать «уклончивый» ответ отчасти снижает роль личностных

характеристик информантов[6], которые в противном случае могли бы привести к снижению достоверности интерпретации полученных результатов. Кроме того, респонденты должны понимать, что возможен ответ «и так, и так», «два варианта»[7] и т. п. на первый и третий (а может быть, и на второй) вопросы.

При этом необходимо понимать, что ответы информантов не следует интерпретировать буквально: если респондент ответил на второй вопрос, что он произносит слово так-то и так-то, из этого вовсе не обязательно следует, что он действительно его так произносит. Здесь могут играть роль и ложные представления о собственной речи (в отношении ударения это случается относительно редко, но для других языковых явлений редкостью не является), и влияние представлений о норме или о «престижности» того или иного варианта[8].

После того как получены ответы, можно переходить к интерпретации результатов. К сожалению, обсуждение того, как на основе этих результатов сделать вывод о региональной норме, находится за пределами данной статьи (некоторые схематичные указания, по необходимости краткие, содержатся в уже цитированной заметке [Шмелев 2004]).

6. Заключительное замечание

В заключение повторим, что методика исследования должна быть адекватна поставленным задачам. Из того неоспоримого факта, что есть задачи, которые нельзя решить иначе, как обращением к корпусу текстов, никак не следует, что такое обращение является оптимальным способом решения любой лингвистической задачи и дело лишь в том, какой корпус выбрать и какой метод его анализа применить. Для многих осмысленных задач наблюдений над реальными текстами недостаточно и необходим лингвистический эксперимент, будь то работа с информантами или интроспекция. Этот вывод мог бы казаться вполне тривиальным, если бы энтузиазм, связанный с использованием корпусных методов иногда не приводил к тому, что его сторонники начинали отвергать методы, не опирающиеся на корпус, и, более того, даже именовать корпусные методы «экспериментальными».

---

[6] Так, стереотипы уверенных, неуверенных и нерефлектирующих носителей языка и влияние указанных характеристик на оценку этими носителями правильности языковых выражений обсуждаются в статье [Беликов 2009].

[7] В принципе вариантов может быть и больше: так, слово *роженица* имеет в узусе три варианта ударения.

[8] Я, напр., на вопрос о том, с каким ударением я произношу слово *йогурт*, отвечу, что ударение ставлю на втором слоге; однако методом самонаблюдения я установил, что на самом деле более чем в половине случаев ставлю ударение на первом слоге.

Литература

1. *Беликов В. И.* Стереотипы в понимании литературной нормы // Стереотипы в языке, коммуникации и культуре. М.: 2009.
2. *Беликов В. И.* Чего не хватает в «оцифрованном мире» лексикографу и социолингвисту // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (2011). Вып. 10 (17). М.: 2011. С. 60–67.
3. *Богданова Н. В., Асиновский А. С., Русакова М. В., Степанова С. Б., Шерстинова Т. Ю.* Звуковой корпус русского языка повседневного общения «Один речевой день»: Концепция и состояние формирования // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (2008). Вып. 7 (14). М.: 2008. С. 488–494.
4. *Шмелев А. Д.* Проблема кодификации в сфере орфоэпии: личный вкус или объективные данные? // Культура русской звучащей речи: традиции и современность. Тезисы докладов международной научной конференции (2004). М.: 2004. С. 116–118.
5. *Shmelev A., Shmeleva E.* Russian Botanical Terms: Towards their Lexicographic Description. Proceedings of the 4th International Conference on Meaning-Text Theory (MTT '09), Montreal, 2009, pp. 339–348.

RUSSIAN-ENGLISH THESAURUS ON COMPUTATIONAL LINGUISTICS

Sokolova E. G. (minegot@rambler.ru)
Russian State University for the Humanities, Moscow, Russia

Kononenko I. S. (irina_k@cn.ru)
A. P. Ershov Institute of Informatics Systems SB RAS, Novosibirsk, Russia

This paper summarizes the experience in the construction of Russian-English information retrieval thesaurus on Computational Linguistics (CL). The need for relating thesaurus terms to the subareas of CL and adjacent sciences is substantiated and the hierarchical structure of subareas is discussed. The kinds of information given in the thesaurus term entry are outlined. A number of terminology description issues are discussed with regard to the specific features of the constructed thesaurus such as bilinguality and insufficient development of Russian CL. Terminological problems are analysed using classification parameters.

Key words: thesaurus, computational linguistics, russian-english, term definition, descriptor, relationship

Introduction

We discuss the results of a two-year (2010–2011) project[1] to develop bilingual Russian-English thesaurus on Computational Linguistics (CL). The project presents high interest in practical (lexicographic) and theoretical aspects taking into consideration the absence of more or less representative descriptions of Russian CL terminology. The closest to the topic works available to date are thesaurus by Nikitina (Nikitina, 1978) that is small and out of date and recent thesaurus INION (Smirenskii, 2007). They describe a small number of CL terms and a lot of purely linguistic terms which have no relation to information and text processing. Other limitations of both thesauri are the lack of definitions of terms and monolinguality. Russian term definitions are given, for example, in the Glossary on Artificial Intelligence (AI) (Averkin, 1992), this field being considered as subordinating or intersecting with CL, but common terms often differ in semantics, for example, *syntactic analysis* in CL and in AI. CL, and especially Russian CL, is not mostly methodological science, as opposed to AI[2].

---

[1] The research was supported by Russian Foundation for Humanities, project No 10-04-12108в.

[2] This also explains the fact that we include subarea of theoretical CL — Generative Grammar — which is related to syntactic analysis task, syntactic annotation task, the task of modelling word order in NLG, etc., and don't include "formal semantics" which corresponds rather to methodological means. See also sec. 2 below.

CL as a research area came into being after the appearance of computers and immediately began reconsidering current descriptions of natural languages (NL) for the purpose of creation of machine translation (MT) systems. Since then CL has extended to other applied areas — Information Retrieval (IR), Information Extraction (IE), etc., and now it remains a new research field whose changes accompany the development of technical devices (computers, global network), scientific progress (linguistics, mathematics) and social advance that affects information technologies. CL is an open area whose theoretical field intersects with linguistics and psychology.

Bilinguality in our thesaurus is symmetric in the sense that every Russian term is provided with English equivalent and vice versa. But the meanings of Russian and English terms often diverge or no equivalence is at hand so that we have to resort to translation. Really, Russian and English states of the art in CL are different. The methodological and technological lagging of Russian CL with regard to the world level is observed on the site of the annual "Dialogue" conference which is the only one in Russia to represent the field of CL in more or less full extent. Also, in Russian there are no textbooks or manuals on CL that would present the area in full enough details, without them being too subjective or compiled from English sources.

Our primary task was to represent the **original Russian CL terminology** and map it into the English one. The goal was not to "translate" Russian CL terminology, but to show it and then merge it with the worldwide state-of-the-art CL. Accordingly, for the Russian part of the thesaurus the collection of proceedings presented at the "Dialogue" Conference in 2000–2010[3] was created and analyzed, it being a helpful source of Russian terms in real use. Proceedings of the International conference "Corpus linguistics" and the manual on corpus linguistics (Zakharov, 2005) served as another source of Russian terms in this one of the most significant subareas of CL. But some empirical and technologically advanced subfields such as Speech Technologies were analyzed mainly in the reverse direction because a lot of terms are lacking in Russian, so the English-language terminological sources have been used and their terms and definitions translated into Russian. As for the English part, dictionaries as well as indices and glossaries of textbooks and manuals have been looked for English terms and definitions.

In (Sokolova et al., 2011) we described the initial phase of the development of Russian-English thesaurus on CL and discussed such questions as the choice of candidate sources of terms, the techniques of machine-aided terminology extraction and selection of basic term list.

In this paper we present the result of the project — bilingual Russian-English thesaurus on CL — section 1; in section 2 the scopes of CL subareas are analyzed in the dynamic perspective; section 3 outlines the types of information given in the thesaurus term entry; in section 4 terminological problems are observed in connection with specific features of the thesaurus and their analysis is given in terms of classification parameters.

[3] Proceedings of the International Conference "Dialogue" are available at: http://www.dialog-21.ru/

1. Results of the project

Results of the project are twofold: a) a list of terms and their descriptions that are usually given in thesauri to represent research area terminology, for example in the above AI thesaurus; b) structuring of CL into subfields and relating terms to subareas or adjacent areas. So the main results of the project are:

- **Russian-English Thesaurus on CL** (**REThes-CL**) exists and is available at: **http://uniserv.iis.nsk.su/thes/**; its content — the list of terms, their definitions, and relations — is open to discussion;
- the content presenting 1671 terms (1031descriptors) on different directions of CL is representative enough to be the base for the thesaurus technology;
- description of CL area organization in the context of adjacent sciences that is not usual in standard thesauri;
- instruction for the term description presenting the specific technology of REThes-CL;
- experience drawn while constructing the thesaurus on CL, which can be characterized as a "scientific — practical area"[4] rather than a "science".

In the next sections we consider the last three points in more details.

2. CL area, its subareas and adjacent sciences

At the initial stage of the project an important problem was missed that is rooted in the abundance of research directions and interdisciplinary nature of CL.

We see that definitions of CL are very vague and in a sense "negative" since they describe CL as "something existing BETWEEN or INTERSECTION of Linguistics, Computer Science and AI (see the term *computational linguistics* in REThes-CL). They don't show the CL true object — processing information presented via NL. Text and speech forms of NL communication plus diversity of directions of text processing divide the CL area into subareas. On the other hand, we can't ignore the adjacent sciences which intersect with the CL proper and form the context for the entire area of CL. The resultant hierarchy of CL subareas and adjacent sciences together with terms for them (top terms) are presented on the website of REThes-CL.

The extensiveness of CL and diversity of its subareas lead to incomprehensibility of certain terms for the users. The parenthetical qualifier is sometimes used to give reference to a subarea, e. g. *pattern (*in *information extraction)*. But the qualifier's main function is to remove the ambiguity of terms, so it should be used only in case of homonyms. The problem of homonymy existing between terms from different subareas of CL (or from CL and adjacent areas) is resolved by introduction of explicit relationship between a term and subarea(s). This ensures for a fully coherent picture of the subarea terminology, which becomes easily accessible for the interested user, on the

[4] Term of O. F. Krivnova

one hand, and could be useful in the indexing purposes to allow for a more precise/accurate search, on the other.

Multiple relationship to subarea or adjacent science is also important due to the fact that subareas are changing over time, so new relations of a term can be added. There are two opposite tendencies of change: generalization — when a separate science forms, and derivation — for child subarea. Both tendencies are observed in IR subarea of CL, historically second after MT. IR is specified in the following core definition: "as an academic field of study, IR … is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections" (Manning et al., 2008). Now IR acquires features of separate science: there appear manuals, courses in the universities, terminological dictionaries, e. g. "SE/SEO glossary"[5].

On the other hand, the traditional subtasks of IR involve structuring and rearrangement of selective information, be it explicit or implicit in the documents: classification/clustering, summarization, IE. In the course of time, however, the development of techniques and tools has led to the establishment of separate applications in this area, independent on the IR task proper. This illustrates the second tendency which leads to coexistence of broad and narrow interpretations of these subareas and their top terms. Thus, in (Grishman, 1997) two definitions of IE subarea are given: 1) a broad view of IE: any method of filtering information from large volumes of text. This includes the retrieval of documents from collections and the tagging of particular terms in text; 2) a narrower definition: the identification of instances of a particular class of events or relationships in a NL text, and the extraction of the relevant arguments of the event or relationship. But in the modern documentation for Gate (Cunningham et al., 2011) this subarea is interpreted as strictly separate from IR: IE differs from IR and traditional techniques in that it does not recover from a collection a subset of documents. Instead, the goal is to extract from the documents salient facts about prespecified types of events, entities or relationships.

An interesting result of the progress over time is the division of CL, which used to be a purely applied science, into Theoretical CL and Applied CL that is made by Hans Uszkoreit (Uszkoreit, 2000). This division is absolutely necessary to differentiate between Applications and models imitating elements of processing speech/texts in human mind (functions of "black box" in cybernetics). Division into subareas in Theoretical CL follows the principle used in Applied CL — we define subarea when there are technologies to resolve **a task**, not following any level-based "theories" of NL in CL, which are very different. We declare the subarea of Syntactic Analysis because there are technologies for the task including dependency filtering method, unification based syntactic-semantic approaches, and others, but there are only experimental attempts in Semantic Analysis. In spite of many experimental or theoretical attempts in this area, especially in the 70–90-es, including R. Shank and many others, to-day the only type of "purely" semantic representations available in the Internet (at least in Russian CL) are "graph structures" in terms of semantic analysis of N. N. Leont'eva (Leont'eva, 2006) available at www.aot.ru. And this approach is close to the tasks of IE.

[5] available at: http://www.infonew.ru/seo_glossary.php#39

3. Representation of terms in the thesaurus entry

The main information items of REThes-CL are domain terms. They are represented by two types of lexical units: single words (mainly nouns) and nominal phrases. The variety of terms is broken down into descriptors (preferred terms) and non-descriptors (variant terms that include synonyms, lexical variants, quasi-synonyms, and abbreviations). The descriptor is chosen as a representative of equivalence class of terms that refer to the same concept and thus exhibit an equivalence relationship.

Terms of any type are provided with ***term name***, ***language***, ***author***, and ***comments***.

The relevance of any term is evidenced by relating it to terminological ***source*** (documents or text collections), this being an additional authority information concerning terms. Where terms or definitions have been extracted from specific available references (glossary, subject matter index, or text), the reference name(s) are given within the term entry. For the collection type sources the ***frequency of occurrence*** is specified.

For the descriptor terms additional attributes are as follows: ***term definition***, ***subarea***, and ***qualifier*** (which is a part of descriptor). The qualifier may refer the term meaning to conceptual category or subject domain, e. g. *accent (pronunciation)* and *accent (prosody), token (corpus linguistics)* and *token (informatics), разметка текста(процесс)*= 'text tagging' and *разметка текста (объект)*= 'tagged text'. Term definitions are not standard for information retrieval thesauri, but it makes the thesaurus a good source of CL knowledge. Definitions are mostly drawn from existing glossaries, papers and manuals, and references to the sources are given. Sometimes one term can have two or three definitions taken from different sources. The list of major sources for terms and definitions is presented in (Sokolova et al., 2011).

The REThes-CL entry presents information concerning interrelations between terms, with following basic types: the above mentioned ***equivalence***, ***hierarchical***, and ***associative***.

Specialized equivalence relationships cover different types of correspondence between descriptors and non-descriptors. ***Synonymy*** holds if a descriptor is substituted for a non-descriptor in all contexts: e. g. *valency* replaces *valence* and *subcategorization; semantic role* substitutes for *semantic case, thematic role, deep case, case role, theta role,* etc. ***Alternative synonymy*** or ***combination synonymy*** are used to replace a non-descriptor by multiple preferred terms with a relation OR or AND in between. For example, *topic* (='a particular subject that the text discusses') and *patient* (='the semantic role of an entity that is not the agent but is directly involved in or affected by the happening') are alternatively used to replace the ambiguous non-descriptor term *theme*. The non-descriptor term *statistical machine translation system* is replaced by a combination of descriptors *machine translation system* and *statistical machine translation*.

Hierarchical and associative relationships hold between the descriptors of each monolingual part of REThes-CL. The ***Hierarchical*** relationship is a connection between broader and narrower terms in the thesaurus. ***Broader*** and ***Narrower***

relationships are further differentiated using three subtypes: ***Generic***, ***Instance***, and ***Partitive***.

A set of descriptors subordinated to the same immediate broader term may be grouped under a ***Hierarchy note*** which specifies a characteristic of hierarchical division of the broader term: *machine translation* is subdivided into *example-based machine translation*, *rule-based machine translation*, and *statistical machine translation* by 'approach' and into *fully automatic translation* and *machine aided translation* — by 'degree of human involvement'.

Associative relationship denotes any nonhierarchical semantic relationship that holds between the descriptors referring to closely related concepts (action/product of action, cause/result, concept/property, agent/counter-agent, etc.).

Translation equivalence relationship connects the descriptors in different languages that refer to the same concepts.

4. Terminological problems arising from immaturity of the Russian CL and bilinguality of thesaurus

Our work to construct REThes-CL has highlighted the effects of immaturity of Russian CL and reflected the unbalanced development of Russian and English CLs, which is the heritage of separateness in the past.

The number of non-preferred terms (synonyms) is small in thesauri for established terminology systems. Our experience with REThes-CL shows another picture: the balance between descriptors and non-descriptors is roughly 2/1 (total descriptors — 1031 and total non-descriptors — 640). As a result of lagging of the Russian CL as compared to the English CL the 'no equivalence' situation, i. e. the absence of Russian terms, frequently occurs.

To analyze and solve the problems with terms in the process of bilingual thesaurus construction the consideration of certain relevant features seems to be helpful. They are enumerated below in relation to the Russian CL:

A. Novelty of a concept — {"+A": new concept, "–A": existing concept};
B. Existence of one, several or zero term name(s) for a concept — {"+B": one term, "++B": several terms, "–B": no term = a new concept borrowed};
C. Existence of stable relation between a concept and a term name (in case of a new concept the relation may be unstable, that results in coexistence of several term variants none of which is preferred by the scientific community) — {"+C": stable term, "–C": unstable term};
D. Potential for the hierarchical semantic change (shift) in the meaning of a term (change from the lower level of the generic or partitive hierarchy to the immediate higher level, or vice versa) — {"+D": possible shift, "–D": impossible shift}.

These parameters are mutually dependent and not all the combinations of their values are real: e. g. <+B, –C> and <– A, – B > are impossible. Other important combinations are presented below.

<(+A|-A), ++B, +C, (+D|-D)> — the descriptor choice problem. The choice is made on base of statistics: traditionally accepted term is selected for indexing purposes, e. g. *актант*, whose frequency of occurrence[6] is 733, is preferred to *аргумент* (616); *валентная структура* (20) is preferred to variants *валентная рамка* (14), *валентностная структура* (3), *схема валентностей* (3).

<+A, ++B, –C, (+D|-D)> — the term for a new concept choice problem. Besides statistics, the problem solving additionally requires the linguist/expert knowledge or intuition, e. g. for the *translation memory* some experts would prefer *архив переводов* (1) in spite of the higher frequency of *переводческая память* (8) and broad use of the calque *память переводов* (0) by translators-practitioners.

<+A, –B, –C, (+D|-D)> — 'no equivalence' in Russian CL and the necessity to coin the term, e. g. the term *целевой фрейм* found in (Kormalev, 2004) as a good match for *template* (used in the IE field to denote the final, tabular output format of IE) and *автоматический перевод устной речи* as a match for *spoken language machine translation*. Note that a lot of terms relating to Speech Technologies have been coined by the entry authors.

<(+A|-A), (+B|++B|-B), +C, +D> — the problem of semantic shift in cases of *модель управления* and *валентная структура* in their different (broad and narrow) meanings. Real use cases of the term *валентная структура* show the ambiguity of this term (quite similar to that of *модель управления)*, as they have both narrow interpretation (syntactic valencies of the predicate word) and broader interpretation (correspondence between semantic valencies and syntactic valencies). The problem is solved by introduction of the qualifier in the term name: *валентная структура* (without qualifier) and *валентная структура (синтаксис)*.

We consider the numeration of possible situations to be a convenient technique to monitor and solve terminological problems in the bilingual thesaurus construction.

Conclusion. Applicability of thesaurus

The thesaurus is a model of the research field in progress, not a description of fixed domain. The developers of REThes-CL face at least two interrelated problems: the bilingual character of the study and the immaturity of the research field. The thesaurus construction helps understand the current state of the art in Russian CL and speed up progress in this area.

REThes-CL is currently used in teaching CL to undergraduate students. It serves as a guide for the students to study the CL structure and state, to familiarize with terms actually in use in the field and to conduct real lexicographic research by describing terms.

Thesaurus will be effective for more accurate definition of basic notions of Russian CL and support mutual integration of CL studies in Russia and abroad; also it will be a good help to the researchers and translators of the field-related literature.

[6] Here the frequencies of occurrence in the Dialogue collection are given, and 'zero' means that the term variant has been found elsewhere but not in the Dialogue texts.

REThes-CL can be characterized as “monitor”, i. e. permanently enlarged under supervision. The opinion is relevant to it: “Like a taxonomy, a thesaurus is never "finished." New findings, and reinterpretation and restating of what is already known, require that terms be added, changed, and occasionally deleted. Continued usefulness of the thesaurus requires an ongoing commitment to updating.” (Milstead, 1998).

On the other hand, we hope that our thesaurus will be useful to solve library catalogs’ problems in describing CL, it being a new scientific-practical area of research. Information in REThes-CL is organized in accordance with standards for information retrieval thesauri, so it may serve as the authority base for indexing and content retrieval of CL texts.

References

1. *Averkin A. N., Gaaze-Rapoport M. G., Pospelov D. A.* (1992). Tolkovyi slovar' po iskusstvennomu intellektu [Glossary on Artificial Intelligence].Moscow, Radio and Communication, 256 p. Available at: (http://www.raai.org/library/tolk/aivoc.html)
2. *Cunningham H., Hainard D., Bontcheva K.* (2011). The Gate Uzer Guide: Text Processing with GATE (Version 6). University of Sheffield, Department of Computer Science. Available at: http://gate.ac.uk/
3. *Grishman R.* (1997). Information extraction: Techniques and challenges. In: Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, International Summer School, (SCIE-97), Springer-Verlag, pp. 10–27.
4. *Kormalev D. A.* (2004). Application of Machine learning to Text Analysis [Prilozheniia metodov mashinnogo obucheniia v zadachakh analiza teksta]. Trudy mezhdunarodnoj konferencii «Programmnye sistemy: teorija i prilozhenija» (Proceedings of International Conference “Programm Systems: Theory and Applications”), IPS RAN, Pereslavl'-Zalesskii, Moscow, Vol.2, pp. 35–48.
5. *Leont’eva N.N.* (2006). Avtomaticheskoe ponimanie teksta: sistemy, modeli, resursy [Automatic Text Understanding: Systems, Models, Resources]. Moscow, Akademiia, 304 p.
6. *Manning C. D., Raghavan P., and Schütze H.* (2008). Introduction to Information Retrieval, Cambridge University Press.
7. *Milstead J. L.* (1998). NISO Z39.19: Standard for Structure and Organization of Information Retrieval Thesauri. Paper presented at the Taxonomic Authority Files Workshop, Washington, DC. Available at: http://www.bayside-indexing.com/Milstead/z39.htm
8. *Nikitina S. E.* (1978). Tezaurus po teoreticheskoi i prikladnoi lingvistike. [Thesaurus on Theoretical and Applied Linguistics]. Moscow, Nauka.
9. *Smirenskii V. B.* (2007). Informatsionno-poiskovyi tezaurus INION po iazykoznaniiu. [Information retrieval Thesaurus on Linguistics by INION]. Moscow, INION, 200 p.

10. *Sokolova E. G., Semenova S. Yu., Zagorulko Yu.A., Kononenko I. S., Zakharov V. P., Krivnova O. F.* (2011). Selection and Preparation of Terms for Russian-English Thesaurus on Computational Linguistics. [Osobennosti podgotovki terminov dlia russko-angliiskogo tezaurusa po komp'iuternoi lingvistike] Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialogue 2011" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2011"). Moscow, pp. 644–655.
11. *Uszkoreit H.* (2000). What is Computational Linguistics? Available at: http://www.coli.uni-saarland.de/~hansu/what_is_cl.html
12. *Zakharov* (2005). Korpushaia lingvistika [Corpus Linguistics]. Saint Petersburg, 2005, 48 p.

AUTOMATIC GENERATION OF TEXT CORPORA FOR CREATING VOICE DATABASES IN A RUSSIAN TEXT-TO-SPEECH SYSTEM

Solomennik A. I. (solomennik-a@speechpro.com)
Speech Technology Ltd., Minsk, Belarus

Chistikov P. G. (chistikov@speechpro.com)
Speech Technology Center Ltd, St .Petersburg, Russia

This paper deals with the problem of speech database design for the needs of unit selection text-to-speech synthesis. An obligatory condition for the naturalness and intelligibility of synthesized speech is a high quality speech database. We propose a computer program developed specifically for the Russian language which creates a phonetically balanced text corpus of given size. We present a description of the program and a comparison of an automatically constructed corpus and some arbitrary corpora. The automatic text corpus generation program is part of a new voice building system for VitalVoice Russian TTS. It helps to supplement a text corpus with missing phonetic units. Further possible improvements of the algorithm are also discussed. We consider several ways to take into account intonational variation of units in a database at the stage of the preparation of a text corpus.

Key words: speech synthesis, unit selection, text corpus, diphone

1. Introduction

Unit selection synthesis is based on determining the best sequence of candidate units from the database. So the first step in developing a new synthetic voice is recording a speech database. Apart from the quality of linguistic processing or selection and modification algorithms, the naturalness and intelligibility of synthesized speech primarily depend on the quality of the speech database. For this reason, various possible units required to synthesize sentences of a natural language must be represented in the database.

There are several ways to deal with the situation when a required unit is missing in the database: simply skipping it; trying to replace it with one having similar characteristics (with possible modification); constructing it from smaller units, etc. But usually it causes a significant loss in naturalness in a particular place where an appropriate unit cannot be found.

The most common way to have all required units is to have large speech databases, representing dozens of hours of speech [1]. But only having a large amount of recorded speech is not enough, it should also be phonetically balanced, i. e. it should, if possible, contain all required units in all possible contexts with various possible

characteristics such as acoustic parameters like fundamental frequency (pitch), duration, position in the syllable, and neighboring phones.

But since the creation of the database requires segmentation, which usually requires at least some manual correction after automatic segmentation ("forced alignment"), the size of the database influences the time necessary to prepare it for use. Besides, large databases are inconvenient to store and search in. So there should be a balance between the size and representativeness of a database.

Usually unit selection speech databases for high-quality TTS contain about 10 hours of recorded speech. Our experiments showed that 2–3 hours are enough for rather satisfactory quality and intelligibility. Less than one hour of recorded speech is insufficient because the necessary variability in such characteristics as pitch and duration will not be achieved, even if all possible units are present.

There are a number of investigations on the automatic corpora construction for different languages, including English, French, Chinese and other [2–5, 11–13]. But for the Russian language our work seems to be the first. The main advantage of our method is that it provides a convenient way for automatic voice creation by the possibility of not simple text selection from a large text corpus, but allows choosing a synthesis unit and creating and editing different preset corpora before selection.

2. Methods and Results

2.1. Instruments

Automatic text corpus generation software is part of the voice building system [8] for VitalVoice TTS [6] developed at Speech Technology Center. A new diphone level was added to database segmentation [7], and at present the TTS engine selects a sequence of diphones for the target utterance. So the examples given in this paper refer to a text corpus for a diphone database, however our program allows obtaining statistics and creating corpora for allophone (i. e. triphone — allophone in context) units as well.

The automatic text corpus generation software was developed on the basis of a program performing the analysis of phonetic unit frequency [9].

Automatic transcription of texts is performed by a modified TTS engine with a disabled automatic break assignment module. Pauses are placed strictly at punctuation marks. This is important because the transcription depends on the positions of pauses. Therefore we imply that the speaker who is being recorded will make pauses in a similar way (it is necessary to control it during recording). It is more convenient than marking pauses in the text with special signs.

Our transcription system for Russian distinguishes 59 types of monophones: 19 vowels and 40 consonants (vowel position in relation to the stressed syllable is specified, and there are 4 additional voiced allophones for word boundaries). Consequently, the number of all possible diphones is 60*60–1 = 3599 (including the pause

context). The number of all possible diphones permitted by the transcription rules (including nonsense words and combinations) is 2335. If various exceptions (such as “боа”) with unstressed /o/ and /e/ are taken into account the number increases to 2759. But not all these combinations are possible in natural human language. For example, only 2176 (from 2759 possible) combinations were found in a large text corpus containing over 460 000 word-forms [10].

2.2. Auxiliary corpora

In order to facilitate the text corpus design, the process is divided into three stages. They are built into the voice building system VoiceConstructor (Fig. 1), which includes all the stages up to installing the new voice.

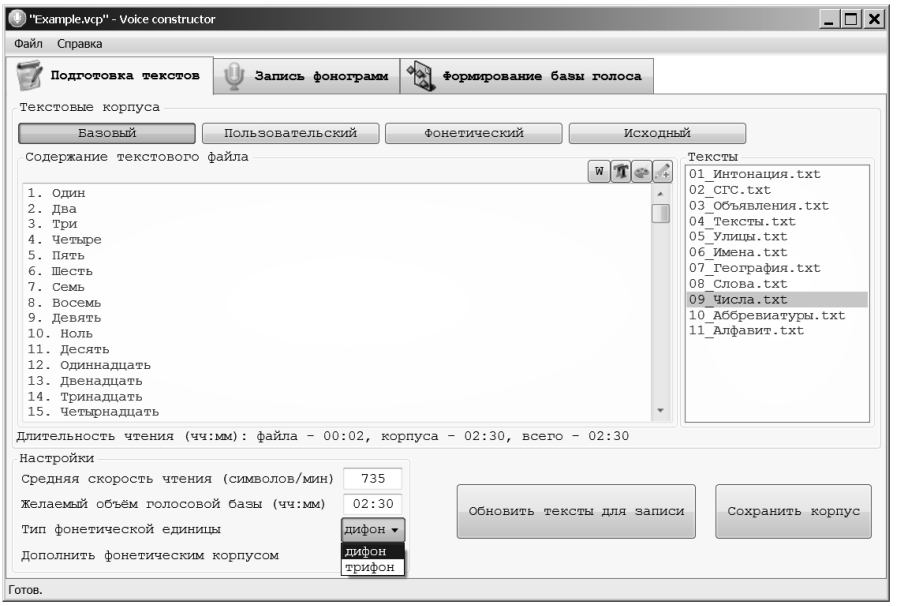


Fig. 1. The window of the basic text corpus in VoiceConstructor

Work with the system begins with specifying the settings of the future sound database. The user must choose the type of the basic unit: diphone or allophone, set the average speech rate (it can also be measured automatically during test recording) and the desired size of the sound database. The program shows the current size of the future database, durations of the current text and the current corpus. The current text is shown in the main window, a list of texts is on the right. The user can edit and delete each one of the texts.

There are four types of corpora: 3 (possible) parts of the resulting text corpus and one large text corpus for statistics and additional phrases:

1. Basic corpus
2. User corpus
3. Phonetic corpus
4. Large source corpus

Basic corpus

The basic corpus is the minimal corpus recommended for recording. It contains various frequent words and word combinations, texts with geographical and proper names, alphabet, abbreviations and some typical texts for speech synthesis applications (e. g. IVR or news). The total volume is 2.5 hours with an average speech rate. Any text can be disabled depending on the user's purpose.

User corpus

The user can upload any texts to the special user corpus. Usually these are phrases that the synthesizer is expected to be able to speak with high quality, e. g. welcome words or terms of use.

Phonetic corpus

The phonetic corpus is actually the main part of the resulting text corpus. It is the part generated automatically to include missing units in the database. It is constructed from sentences selected from the source corpus.

Source corpus

At first the source corpus was supposed to be a fixed large text corpus needed to obtain diphone statistics and to select phrases with missing units. But then we decided that it is more useful to allow users to change and upload texts. There are several reasons for it. First, for domain-specific TTS it is very important to choose sentences from a particular domain or genre. Apart from typical frequent words and combinations, we need to take into account possible changes in speech rate, manner and even pitch and timbre (compare reading books and announcements in an airport). Besides, a large corpus requires significant time for its processing. But the source corpus should be sufficient to represent the majority of possible units, and we recommend to include into it at least 10 hours of speech at an average speech rate (about 70 000 word-forms). An example of the source corpus which was compiled from different texts from web resources (news, politics, IT, fiction) is included to the system.

2.3. Algorithm

The algorithm for the generation of the phonetic corpus includes the steps shown in Fig. 2. First, the system transcribes all the necessary texts in a way described above. Then it calculates the desired volume of the phonetic corpus using the data about the total desired corpus size and the size of the basic and user corpora (if any).

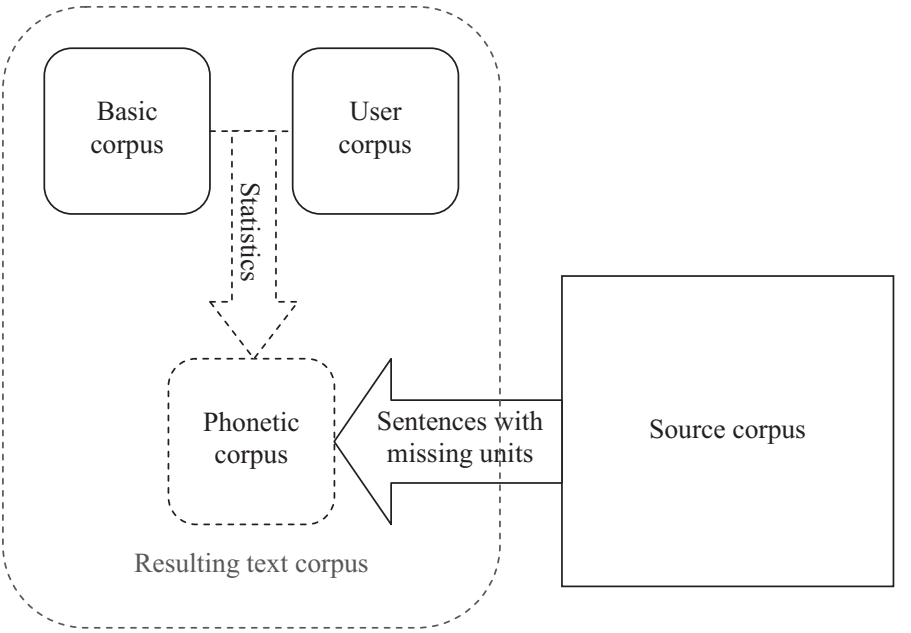


Fig. 2. Generation of text corpus

Sentences are chosen from the source corpus depending on how many missing units they contain, sentences with maximum missing units are taken first. If two sentences contain the same number of missing units, sentence with less frequent diphones will go first. The procedure is the same for infrequent diphones in the case when all diphones from the source corpus are already present in the basic and user corpora. The addition of sentences will stop when the resulting text reaches the desired size. This allows us to obtain a text of minimum size with maximum missing units. Of course the sentences in the text are not connected to each other in meaning, but each separate sentence is not artificial and easy to read. Speech databases usually consist of a number of files with only one sentence in each file, so sentences in VoiceConstructor are also recorded separately. Reading nonsense text also helps in controlling the emotions of the reader, as it is easier to read with constant rate and timbre.

After adding the phonetic corpus, all three corpora can be saved as text files and are displayed at the next stage when sound files are recorded.

2.4. Example of automatic text generation

To demonstrate the work of the program we compared a text constructed by the program to an arbitrary text or texts of the same size. We took relatively small amounts of data in order to speed up data processing. We generated a text for an hour of speech at the average speech rate (about 8000 word-forms) and compared its diphone and triphone statistics with texts of two different styles. The first was part

of the novel “Obmen” (The Exchange) by Soviet writer Y. Trifonov and the second was part of “Dialog 2010” proceedings (only Russian texts). Both also contained about 8000 word-forms.

Source text corpus contained more than 100 000 word-forms. It was drawn from different texts from web resources (politics, IT, fiction).

The statistics were obtained and compared by means of a specially designed program performing the analysis of phonetic unit frequency [9]. The results are shown in the Table 1 below:

Table 1. Diphone and triphone statistics in various texts

Text	Number of word-forms	Number of diphones (missing: in source/ in text)	Number of triphone (missing: in source/ in text)
“Obmen”	7835	1566 (5/462 — 23 %)	9852 (587/16054)
Dialogue 2010	8590	1488 (10/545 — 27 %)	8494 (415/17240)
Generated text	7557	1974 (0/49 — 2 %)	11 704 (0/13615)
Source corpus	106 260	2023	25 319

As we can see from the table, scientific text is statistically the poorest in phonetic variability, although it contains some unique diphones. The automatically generated text (we should note that it is generated in order to obtain diphone coverage, triphones were not taken into account) is the most representative and contains almost all the diphones presented in the source text. The table also demonstrates why diphone units are preferable to allophones: a context of a diphone in most cases is not as important as that of an allophone, so the required number of diphones is many times smaller.

A part of the resulting text corpus is presented below:

(1) *Еще более скептично смотрит на перспективу привлечения силовых структур к банковскому контролю эксперт «ФиэнЭкспертизы» Наталья Борзова: "Никакой потребности в каких-то дополнительных контролирующих структур сейчас нет, было бы куда полезнее, если бы ФСБ и МВД занимались своим делом, например, проверяли бы те адреса, по которым регистрируются фирмы-однодневки, тем более что силовые структуры и так уже имеют полномочия проводить выемку документов и получать всю необходимую для следствия информацию.*

Один из самых больших обманов 1 апреля, о котором долго потом вспоминали газеты и журналы, произошел в Лондоне в 1860 году, когда несколько сотен джентльменов с их чопорными английскими леди получили приглашение прибыть «на ежегодную торжественную церемонию умывания белых львов, которая состоится в Тауэре в 11 часов утра 1 апреля».

Антонио Вальдес-Гарсиа узнал из своей истории болезни, что ему были нанесены следующие телесные повреждения: закрытая черепно-мозговая травма, сотрясение головного мозга, закрытый перелом правой бедренной

кости с отрывом большого вертела, открытый двусторонний перелом нижней челюсти: центральный и суставного отростка слева со смещением, перелом альвеолярного отростка верхней челюсти слева, закрытый оскольчатый внутрисуставный перелом основания ногтевой фаланги 1 пальца левой стопы, множественные ушибленные раны лица, правого коленного сустава, левой стопы.

Президент Квасьневский, заметил Валенса, даже не смог достойно закончить свой президентский срок, помиловав экс-главу МВД и товарища по партии Збигнева Суботку, который был осужден на 3,5 года тюрьмы за передачу польской мафии секретной информации о готовящейся против нее операции (тем самым экс-министр поставил под угрозу жизни полицейских).

Он некоторое время слушал топот шагов убегающего Тигра после того, как тот скрылся за углом, как вдруг уловил звук дизельного двигателя средней мощности и сам припустил бегом по тротуару — как раз вовремя, чтобы увидеть задние габаритные огни грузовика, очень похожего на фургон, развозящий экспресс-почту, но оснащенного параболической антенной на крыше.

«Неформальных» разговоров в Сочи будет более чем достаточно, и, возможно, именно они будут играть ключевую роль, но, чтобы придать саммиту видимые результаты, стороны подпишут два соглашения в рамках одного из «общих пространств», как принято определять сферы практического российско-европейского взаимодействия: об упрощении визового режима для отдельных категорий граждан и о реадмиссии.

Что касается потенциальных «темных лошадок», то при подобном «уплотненном» сценарии, требующем уже на раннем этапе колоссальных затрат на рекламу и другую подготовительную работу в десятках штатов сразу, уверенно чувствовать себя может разве что нынешний нью-йоркский градоначальник мультимиллиардер Майкл Блумберг.

Здесь, правда, надо заметить, что и в прошлые годы страсти накалялись (в меньшей степени, конечно) именно к 16 декабря, поскольку в этот день 14 лет назад Генеральная Ассамблея ООН отменила свою же резолюцию 1975 года, объявлявшую сионизм формой расизма.

3. Conclusion

The program presented in this paper not only generates a phonetically balanced minimal text corpus but gives the user complete freedom in editing and choosing the type and style of the resulting text. It is very important when considering various possible applications of speech synthesis. But there are still some problems to solve.

One of the issues is when our sound database can be considered complete. A presence of only one realization of a particular diphone is usually not enough because it cannot cover all possible characteristics of a diphone such as pitch, energy

or duration without modification. For some units it is not as important as for others (for example, voiceless plosives vs. vowels under phrasal stress).

Natural sounding intonation is significant for high quality speech synthesis. It is very important to have units with at least two types of pitch movement to model phrasal and emphatic stress. The main tone movements normally appear on stressed syllables, so we may reduce the number of required diphones to those with stressed vowels (513 are possible). The intonation on them can be controlled by special signs for the speaker or simply by punctuation (full stop, comma or question mark). There are two possible solutions: automatically selecting sentences with the required punctuation and phonemic structure from a large text corpus, or manually constructing short sentences for the most frequent stressed units. The latter is more preferable since the former leads to a strong redundancy and some contexts (i. e. contexts with question marks) may not be present even in very large text corpora.

References

1. *Black A. W.* Perfect Synthesis for all of the people all of the time // Keynote, IEEE TTS Workshop. Santa Monica, CA, 2002.
2. *Bozkurt B., Ozturk O., Dutoit T.* Text design for TTS speech corpus building using a modified greedy selection. The 8th Eur. Conf. on Speech Communication and Technology (EUROSPEECH), Geneva, Switzerland, pp. 277–280.
3. *Chevelu J., Barbot N., Boeffard O., Delhay A.* Comparing set-covering strategies for optimal corpus design. Proceedings of the 6th International Language Resources and Evaluation, 2008.
4. *Chevelu, J., Barbot N., Boeffard O., Delhay A.* Lagrangian relaxation for optimal corpus design. Proceedings of the 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6). Bonn, Germany, 2007.
5. *Francois H., Boeffard O.*, Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem. Proc. Eurospeech, 2001
6. *Oparin I., Talanov A.* Outline of a New Hybrid Russian TTS System // Proc. of the 12th International conference on Speech and Computer, SPECOM 2007. Moscow, Russia, 2007. Pp. 603–608.
7. *Prodan A. I., Korolkov E. A., Oparin I. V., Talanov A. O.* MULTI-TIER MARKUP OF SPEECH CORPUS FOR HYBRID RUSSIAN TTS SYSTEM “VITALVOICE”. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2009” [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2009”]. Bekasovo, 2009.
8. *Prodan A. I., Talanov A. O., Chistikov P. G.* Voice building system for hybrid Russian TTS system “VitalVoice”. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2011” [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2010”]. Bekasovo, 2010.

9. *Smirnova N., Chistikov P.* Software for automated statistical analysis of phonetic units frequency in Russian texts and its application for speech technology tasks. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2011" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011"]. Bekasovo, 2011, pp.632–643.
10. *Smirnova N., Chistikov P.* Statistics of Russian Monophones and Diphones. Proc. of the 14th International conference "Speech and Computer", SPECOM 2011, Kazan, 2011, pp. 218–223.
11. *van Santen, J. P. H., Buchsbaum A. L.*, Methods for optimal text selection. Proc. of Eurospeech, Rhodes, Greece, 1997, pp. 553–556.
12. *Wei Z., Yayu L., Ye D.* Automatic Construction for a TTS Corpus with Limited Text. Measuring Technology and Mechatronics Automation (ICMTMA), 2010.
13. *Wu H., Xu B., Huang T.* Automatic Corpus Selecting Algorithm Based on Triphone Models, Journal of Software, 2000, 11(2), pp. 271–276.

USING SENTIMENT-ANALYSIS FOR TEXT INFORMATION EXTRACTION

Solov'ev A. N. (a.solovyev@i-teco.ru),
Antonova A. Ju. (a.antonova@i-teco.ru),
Pazel'skaia A. G. (Pazelskaya@i-teco.ru)

«I-Teco», Moscow, Russia

This article aims to demonstrate relationship between sentiment-analysis and information extraction from text. To illustrate the idea we propose a sentiment-based summarization prototype instrument. The performance of the method in comparison with 3 another summarization methods is tested on a small corpus of mass media and blog posts.

Key words: sentiment-analysis, semantics, emotional component of meaning, information, automatic text summarization, information extraction

1. Introduction

Automatic document sense extraction is considered to be among crucial problems of natural language processing. Usually a document contains numerous noun phrases. The extraction of the most characteristic NPs appears to be a difficult task partly solved by statistical methods or with the help of thesauri for a certain domain ([Lukashevich 2011][1]).

Blog texts procession poses even more difficult problems at all levels: lexical, syntactic and stylistic. Commonly a blog post consists of the post itself and a hierarchy of comments. Identifying the antecedent of a comment and the part of the post commented on is often problematic even for a human. To do this, one needs first to identify the topic(s) of the post (or a parent comment), the topic of the comments and look for matches between them.

In this case semantic analysis based on tokenization followed by morphological and syntactic analysis turns to be just helpless because of lexical, syntactic and stylistity specifics of the blogosphere. Polythematics as a typical peculiarity of blogs (unlike news) adds even more confusion to an automatic processing task.

Automatic text summarization aims to extract key information from texts, therefore it can be considered as an approach to text understanding.

Currently automatic summarization widely uses approaches from information theory (e. g. self-information, entropy [Gusev et al. 2005], Bayes models, Markov

[1] Chapter 22: Socio-political thesaurus and automatic annotation, pp. 414–446

chains and nets [Fung et al, 2003] etc.). Despite of the growth of sophistication of algorithms, getting a high-quality summary of a randomly chosen text is still something to be desired. Preliminarily constructed thesaurus of a specified domain can partly solve the problem, but it involves new troubles such as processing speed decrease with the increase of hierarchy complexity and thesaurus volume, as well as domain-specificity, homonymy etc. ([Barzilay R., Elhadad M. 1997]).

In our article we are discussing the semantics of a message focusing on its emotional component. We postulate that any news message or blog post always contains an emotional element. In other words, our starting point is an emotive function of communication as opposed to informational aspect. Our claim is that communicative fragments of a message with highest emotional charge are most important for the speaker and, therefore, most informative for the receiver. (For more details about communication functions see [Iakobson 1975, Iarceva et al. 1990].)

Thus, our approach can be applied is most applicable to the communication sphere that exploits emotive function (first of all, to blogosphere and mass-media). Obviously, good result is not expected where informative function dominates (science publications, law acts, instructions etc.). Still the emotive component is characteristic for texts of most styles and genres. It is almost always present in news messages, social network texts and, certainly, in verbal communication[2]. Hence, we propose the following hypothesis: the information retrieval task can be resolved via sentiment analysis of every sentence in the message and its sentiment force evaluation.

To verify our hypothesis, we used sentiment-based method of automatic text summarization. Our method was further compared to two other summarization methods (based on self-information and on TF/DF measure) and also to the performance of TextAnalyst system (Microsystems).

The structure of the rest of the article is as follows. Section 2 presents sentiment analysis as an instrument for text summarization. After basic concepts of sentiment analysis (2.1) a very brief overview of the method of calculating sentiment degree of a sentence is given (2.2). As in our investigation we used a program for sentiment analysis, some details of its functionality are described in 2.3 and 2.4. Section 3 focuses on the experimental verification of the aforementioned hypothesis. The conditions of the experiment are given in 3.1. As in our experiment we compared our method with three other ones, those are described in 3.3 after a very brief overview of summarization methods (3.2). The results of the experiment are presented in Section 4 and the interpretation and further perspectives conclude the article (5).

[2] "A man, using expressive features to indicate his angry or ironic attitude, conveys ostensible information [...]" (Jakobsón R. Language in Literature. Ed. Krystyna Pomorska and Stephen Rudy. Cambridge (Massachusetts), Belknap Press, 1987, p.67)

2. Sentiment and its use for text summarization

2.1. What is sentiment

We have already discussed the concept of lexical sentiment and the approach to its automatic detection in our previous article [Pazel'skaia, Solov'ev 2011]. We define lexical sentiment as either lexeme-level or communication fragment-level emotional component. The sentiment of a whole sentence is calculated on the basis of lexical (given in a dictionary) sentiment of its words and communicative fragments[3], considering linkage rules. Sentiment is always related to a special object of sentiment. Besides, taking into account the sentiment subject (the author of the statement) brings us to the task of opinion mining[4]. The sentiment score (the force of the sentiment component) is another valuable parameter for analysis. Though in the general case, emotional space can be described as multi-dimensional, we examine one-dimensional sentiment space, the sentiment being measured over "positive-negative" scale together with the force of the positive or negative sentiment expressed — sentiment score.

2.2. Calculating the sentiment

The polarity score is evaluated using contribution of several factors and is calculated within each sentence. The key factor is those adjectives, adverbs and collocations (with either neutral or ambiguous sentiment) that strengthen the emotional component. The sentiment score for these fragments is defined in sentiment dictionaries and has three levels (neutral, strong, very strong). Tonal adjectives, nouns and auxiliary parts of speech have neutral sentiment score. This is a technical simplification, imposed by the fact that these tokens combine with emotion-strengthening components to form sentiment chains. This operation together with word chains sentiment recalculation sometimes causes difficulties, therefore it is easier to reserve lexically defined sentiments strength for adverbs, tonally neutral adjectives, nonverbal collocations, and for predicative elements, namely, for verbs and verb collocations.

Sentence-level sentiment is calculated as a sum of sentiment strength of joined word chains in the sentence. Suppose that after building sentiment chains a sentence is represented as "subject chain" — "predicative chain"— "object chain". Then the system calculates sentiment value (positive/negative) and polarity score for each chain.

---

[3] Communicative fragments are speech fragments of different length that are stored in the speaker's memory as fixed segments of speaker's language experience that he (she) uses uttering or interpreting expressions [Gasparov, 1996]. Usually a communicative fragment is longer than a word and shorter than a sentence.

[4] Demo systems for sentiment analysis and opinion mining developed in CJSC "I-Teco" can be found at http://x-file.su/tm/ (sentiment analysis) and http://x-file.su/ds/ (opinion mining).

After that a special set of rules defines sentiment of the whole sentence, and the sentence polarity score is calculated as a sum of polarity scores of the three chains. Therefore, when calculating polarity score, we do not look at its value, these two parameters are independent.

Text-level sentiment takes into account the number of sentences with expressed sentiment in the text and their polarity score. It should be mentioned that even sentiment bearing sentences may have neutral (or zero-valued) polarity score if sentiment is weakly expressed, i. e. there are only lexis with neutral sentiment strength without any sentiment amplifiers.

2.3. System output

For readability purposes, in the system output the sentences where the sentiment score is higher than neutral are displayed with bigger font, the font size being proportional to the sentiment score. Depending on the sentiment score and sentiment value, a sentence can bear one of the following seven sentiment characteristics: weak negative, medium negative, strong negative, neutral, weak positive, medium positive and strong positive.

Examples (in italics):
В таких условиях состояние здоровья мамы не улучшается, а ухудшается.
(In such circumstances, mum's health is not improving, but is getting even worse.)– mixed mean negative and strong negative sentiment;
Кроме того, *он планирует запретить пропаганду культа личности*, пишет газета “Коммерсантъ”.
(Besides, he plans to prohibit the cult of personality propaganda, writes “Kommersant”.) — neutral and strong positive sentiment (italics);
Независимость автономии на данный момент признали 127 государств.
(127 countries have already recognized independence of the autonomy.)– weak positive sentiment;
Он хорошо освоил программирование, ***но так и не научился писать стихи.***
(He became keen in programming, but haven't learnt to write poems.)– weak positive (italics) and weak negative (bold italics) sentiment.

2.4. Sentiment-score in text summarization

Our automatic sentiment analysis system has an option of extracting non-neutral sentences only. The filter uses absolute values of sentiment score. If a sentence includes sub-sentences with different sentiment score, larger score is chosen. We used this filter function to get an emotive summary of a document. This gave us a small corpus of texts and their summaries which was used to check the hypothesis stated in the Introduction.

3. Experimental verification of the hypothesis

Our experiment was as follows. From each document those sentences were extracted that had absolute sentiment value over zero. The summaries obtained in this way were compared to those received with pure statistical methods (see Section 3.2). Each summary of every document was evaluated by a number of respondents.

3.1. Material and respondents

For current research we took originals of news texts and blog posts (source texts) and also their summaries, i. e. sequences of all non-neutral sentences from these texts. Every source text selected for the experiment meets the following criteria:

or current research we took originals of news texts and blogs and also their summaries, i. e. all chains of non-neutral sentences from these texts. Every selected text answers the criteria:

1) not very short (more than five sentences) and not very long (not more than 2000 symbols);
2) no thematic restrictions were posed for news texts (our sample includes texts on politics, incidents, sports, science, culture etc.), as sources we chose news portals www.rbc.ru, www.lenta.ru, www.rian.ru, www.utro.ru;
3) blog texts were picked from blog corpus belonging to CJSC "I-Teco" (http:// www.i-teco.ru), none of the texts were news reports, advertisements or announcements.

Finally 25 random (with restrictions mentioned) official news texts and 25 blogs posts were taken. The number of texts was restricted by the experiment time limitation (2–3 hours for each respondent, in order to not to exhaust the experiment participants).

19 volunteers (age 20–60) participated in the experiment. None of them was experienced in analyzing summaries.

Our respondents were asked to read the source text and evaluate each summary according to 2 scales, each having three values:

1) summary quality, that we for convenience called precision[5] (good (contains all relevant information) / normal (contains only part of the relevant information) / bad (noise, all the relevant information is lost));
2) redundancy (no redundancy / little redundancy / very redundant).

The redundancy score (though sometimes correlated with the precision score) was taken to provide better evaluation of those summaries that apart from key

[5] In current research, precision is an averaged psycholinguistic score of respondents' subjective estimation of summaries. It appears impossible to estimate precision objectively as no tools for message semantics estimation exist.

sentences contain extraneous information. The participants were not asked to evaluate logical consistency, as it would bring us far from the goal of our investigation.

The respondents were unaware of the summary extracting methods we used. All texts were printed with the same type and color. Titles of source texts were erased. Pauses were allowed while completing the task.

3.2. Experiment preparation: automatic summarization of the documents

Though many different approaches to text summarization were presented in the literature, most of them can be divided into 2 groups. First is so-called quasi-summarization, where summaries consist of sentences extracted from the source document without any modifications. The second group algorithms, summary generating ones, generate a new text that somehow interprets the source one. Our experiment was held within the first approach (sentence extraction; for more detailed overviews see [Han, Mani 2000; Das, Martins, 2007; Ganapathiraju 2002]) Summarization algorithms can be further divided into two types: 1) ones that use statistical information on words and collocations distribution and 2) those based on linguistic information i.e. exploiting tokenization and syntax information together with dictionaries and thesauri.

As linguistic rules are typically domain-dependent and genre-dependent, it appears to be difficult to use them when working with texts belonging (like in our case) to different domains and genres, especially with blog posts. Therefore, only statistical algorithms were taken to be compared with our system. These are 1) an algorithm based on TF/DF; 2) an algorithm exploiting self-information measure. The TF/DF measure is widely used in summarization systems, e.g. in HMM (see [Fung 2003]). Second method was used to demonstrate the difference between the concept of self-information that was brought to linguistics from information theory [Shannon 1963] and the semantic notion of information. We also used for comparison output of TextAnalyst NLP system, a product of Microsystems company.

3.2.1. Automatic summarization using self-information[6] measure

This algorithm can be briefly described as follows. First in our mixed corpus (blogs post and mass-media, about 100 million words) all the personal names were replaced with a special tag. Second the corpus was statistically processed with free distributed SRILM-tools[7]. Then for each unigram of the language model a logarithm of its probability (i.e. its amount of self-information) was calculated. The possibility to come across a new word was also considered (Katz's back-off model). Finally from each source text the system extracted 1/3 of its sentences (with maximum value

---

[6] Self-information is a measure of the information content associated with the event with certain possibility. For an independent event X with possibility P Hartley formula was used: $I(X) = -\log_2 P(x)$.

[7] Available at http://www.speech.sri.com/projects/srilm

of self-information averaged and normalized by the number of unigrams in a sentence). This extraction was taken as a summary.

3.2.2. Automatic summarization using TF/DF measure

The TF/DF based algorithm [Salton 1988] on step 1 evaluated weights of terms (unigrams, bigrams and trigrams) in each sentence (every n-gram was considered as a separate term). We cut off rare n-grams and took only 50 % most frequent unigrams, 15 % bigrams and 5 % trigrams. Weights for news and blogs were evaluated separately. On step 2, by intersecting weighted term-document matrix with all terms in a document, normalized sum of terms for every sentence was calculated. Then (step 3) weights were normalized by the sentence length. The summary included 1/3 of document sentences with the highest normalized term weight.

3.2.3. Automatic summarization by TextAnalyst

To compare to summaries extracted by our system, we used commercial program TextAnalyst[8] that provides a summarization module. The system uses neural network model for inner text representation and ranks sentences using weights for previously identified key terms. With this instrument we collected summaries for news texts (about 80 % texts were successfully processed). As the program failed with most of the documents from social networks, TextAnalyst was excluded from comparative analysis of blog summaries.

3.2.4. Experiment conditions

Neither statistical algorithms nor the sentiment-based one had the advantage of using syntactical analysis. The weight of terms from the first sentence was not increased, although it could have been quite reasonable for news texts. No special heuristics for maximizing summarization accuracy were used (such as headlines contribution — technique that often turns to be absolutely useless when dealing with blogs).

Finding the best algorithm was not our aim. Our purpose was to show that sentiment-based method possesses its own competitive advantages and can be used either separately or within some complex method. So we emphasize that (except for TextAnalyst system) each method is not a ready to use one and can be equally improved. Thus, the three summarization methods for news texts and 2 methods for blogs were used in our experiment to compare with our sentiment-based method.

4. Results

Having received the results from our respondents, we compared the scores for each method. Tab.1 below presents this comparison.

[8] Free version at http://www.analyst.ru/index.php?lang=eng&dir=content/products/&id=ta

Table 1. Average precision and redundancy values and their standard deviation for the summarization methods. For the three-point scale where 1 is "bad" and 3 is "good"

	News		Blogs		News		Blogs	
Method	**Precision**	**Stand. deviation**	**Precision**	**Stand. deviation**	**Redundancy**	**Stand. deviation**	**Redundancy**	**Stand. deviation**
Self-information	1.69	0.22	1.71	0.24	1.88	0.33	1.94	0.36
TextAnalyst	1.79	0.21	–	–	**2.41**	0.23	–	–
Sentiment	**2.38**	0.25	**2.07**	0.27	2.08	0.28	**2.24**	0.25
TF/DF	2,04	0,23	1,91	0,20	1,99	0,25	1,74	0,27

Below the results are illustrated with histograms (see Figure 1).

The examination of processing results for both blogs and mass-media texts showed that the precision value was a bit better for the summarizing method extracting most expressive sentences from texts. (Although the advantage 2.07 against 1.97 (for blogs) can be considered as statistically insignificant, we would remind that our main purpose was not to prove our method highest accuracy compared with another methodologies.)

The widely used methodology based on TF/DF measure showed the second best result for precision test. Next is TextAnalyst that also had the best value for redundancy scale (on news texts).

The worst scores for both news and blogs categories were obtained by the methodology of extracting sentences with highest self-information values. Such comparatively low result had been anticipated. In information theory, the less probable event is considered to be most informative. Probabilistic concepts from information theory can be successfully applied for extracting statistically significant lexical, grammatical or syntactical templates (e. g. in collocation extraction, document flow frequency analysis etc.). Nevertheless, the result of application self-information measure to semantic analysis is very poor because the concept of "information" in mathematical theory of communication differs from the one in natural language theory.

The average precision value for mass-media is better than for blogosphere (2.04 against 1.89 respectively). Blog discourse structure is much more complicated, due to indirectly expressed sense (that often demands context understanding and further including of additional knowledge), polythematic character of a single post, low narration consistency.

The minimum redundancy value is shown by mass-media summaries by TextAnalyst system. Sentiment-based methodology has the second rank. Both statistics-based summaries have relatively low and quite similar redundancy because those algorithms always extracted 1/3 of source text.

We scored the degree of unanimity of our respondents with non-parametric Kendall's coefficient of concordance. This statistic was chosen because it makes

no assumptions regarding the nature of probability distribution. We received the following results. Mass-media texts were scored with coefficient of W=0.49 for precision and W=0.29 for redundancy. The coefficient for blogs is lower W= 0.36 for precision and W=0.22 for redundancy. The redundancy criterion was scored with higher variability level. We suppose that this happened because respondents interpreted the criterion differently.

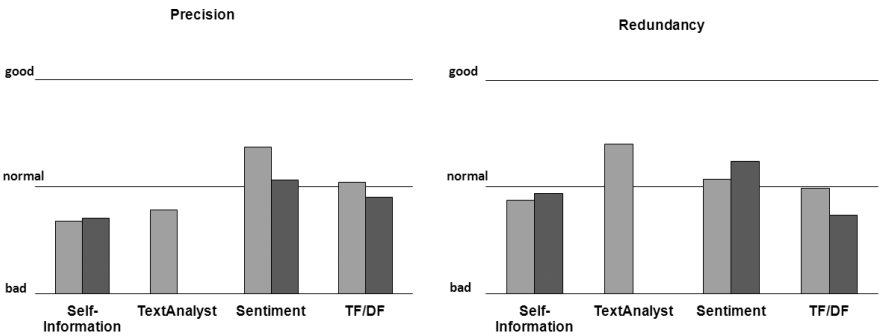


Figure 1. Comparison of four summarization methods. The left picture shows precision and the right one redundancy criterion. Light-grey boxes are used for media and dark-grey for blog texts

5. Conclusion and perspectives

This paper presents the following results. We applied sentiment analysis method to the task of summarization. This approach proved to be competitive with other previously developed summarization methodologies. The specific of our method lies in its ability to get quite a good summary without prior statistical processing of large document collections. Nevertheless, the contribution of statistical features will improve the summarizing accuracy. Application of coreference links and anaphora resolution would also significantly raise the quality of summaries.

Another contribution of our research is the proof (though indirect) of independence between document semantic characteristics and the amount of self-information from the mathematical theory of communication.

A possible application of the proposed method is the "text emotional temperature" estimation in its dynamic as well as comparing sentiment force for different sources commenting the same event. The theoretical issue of our research is that at least for some types of texts on emotional function of communication is inseparable from information function, emotive and informative peaks of text coinciding. In general case it might be untrue, but still this generalization holds for those texts people deal with in their everyday life. Thus, emotionality is the essential characteristic of interpersonal communication.

Acknowledgement

We thank all the respondents who took part in our research and we would like to acknowledge the help of Denis Mazur who had designed the architecture of sentiment-analysis instrument.

References

1. *Barzilay R., Elhadad M.* 1997. Using Lexical Chains for Text Summarization. — ACL/EACL Workshop Intelligent Scalable Text Summarization.- Madrid.Das D, Martins A. F.T. A Survey on Automatic Text Summarization.Literature Survey for the Language and Statistics II course at CMU, November, 2007
2. *Braslavski P., Kolychev I.* eXtragon: an Experimental System for Automatic QueryBiased Summarization of Web documents [eXtragon: eksperimental'naia sistema dlia avtomaticheskogo referirovaniia veb-documentov] ROMIP-2005, St-Petersburg, Russia, 2005, pp. 40–53.
3. *Das, D., Martins, A. F. T.*: "A Survey on Automatic Text Summarization"; Literature Survey for the Language and Statistics II course at CMU (2007)
4. *Fung P., Ngai G., Cheung C.-S.,* Combining optimal clustering and hidden Markov models for extractive summarization, in: ACL Workshop on Multilingual Summarization and Question Answering, Association for Computational Linguistics, Morristown, NJ, USA, 2003, pp. 21–28
5. *Ganapathiraju M. K.* "Relevance of Cluster size in MMR based Summarizer: A Report 11-742: Self-paced lab in Information Retrieval". November 26, 2002.
6. *Gasparov B. M.* (1996) Aizyk, pamiat', obraz. Lingvistika iazykovogo sushchestvovaviia [Language, memory, image. Linguistics language of existence]. Moscow, Novoe literaturnoe obozrenie [New literary review].
7. *Gusev V. D., Miroshnichsenko L. A., Salomatina N. V.* Thematic Analysis and Quasi-Abstracting of Text Using Scan Statistics [Tematicheskii analiz I kvasireferirovanie teksta s ispol'zovaniem skaniruiushikh statistic]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2011" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2005"]. Zvenigorod, 2005, pp. 121–125.
8. *Hahn U., Mani I.* The Challenges of Automatic Summarization, IEEE Computer, November 2000, pp. 29–36.
9. *Iakobson R. O.* (1975) Lingvistika i pojetika: Strukturalizm: «za» i «protiv». [Linguistics and Poetics: structuralism "pro" and "contra"]. Moscow.
10. *Iarceva V. N., Klimov G. A., Zhuravljov V. K.* Sovetskoe jazykoznanie // Lingvisticheskij jenciklopedicheskij slovar' [Soviet Linguistics // Linguistic encyclopedic dictionary]. Jarceva (ed) — Moscow: Soviet encyclopedia, 1990.
11. *Lukashevich N.* Tezaurusy v zadachakh informatzionnogo poiska [Thesauri in Information Retrieval Tasks]. Moscow, MSU, 2011, pp.265–271 and pp.414–446.
12. *Pazel'skaia A., Solov'ev A.* A Method of Sentiment analysis of Russian Text [Metod opredeleniia emotzii v tekstah na russkom iazyke]. Komp'iuternaia Lingvistika

i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2011" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011"]. Bekasovo, 2011, pp. 510–523.
13. *Salton G. and Buckley C.* (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24 (5), pp. 513–523.
14. *Shannon C. E., Weaver W.* The Mathematical Theory of Communication. Univ. of Illinois Press, 1963.

СОЮЗЫ, КОННЕКТОРЫ И ТЕОРИЯ ВАЛЕНТНОСТЕЙ[1]

Урысон Е. В. (uryson@gmail.com)

Институт русского языка им. В. В. Виноградова РАН, Москва

Объект работы — союзы и коннекторы, т. е. слова, которые маркируют связи между т. н. сверхфразовыми единствами. Коннекторы, подобно союзам, имеют как минимум два семантических актанта, однако поверхностное выражение этих актантов не может быть описано строгими синтаксическими правилами. Вводятся определения для описания специфики их выражения. Демонстрируется, что некоторые союзы, как сочинительные, так и подчинительные, могут функционировать как коннекторы, а другие — не могут, причем данная информация должна подаваться в словаре.[2]

Ключевые слова: союз, коннектор, семантический актант, синтаксическая валентность, невыраженная валентность, сфера действия, предложение, текст.

1. Объект исследования и постановка задачи

«Союз — это служебная часть речи, при помощи которой оформляется связь между частями сложного предложения, между отдельными предложениями в тексте, а также (это относится лишь к некоторым союзам) связь между словоформами в составе простого предложения» [Грамматика-80: 713]. Примеры:

(1) *Я встану,* ***когда*** *будет светло.*

(2) *Стало светло,* ***и*** *я встал.*

(3) *Стало светло.* ***И*** *я встал.*

---

[1] «Работа выполнена при финансовой поддержке Программы фундаментальных исследований отделения историко-филологических наук РАН «Язык и литература в контексте культурной динамики», гранта РГНФ No. 10-04-00273а и гранта НШ-6577.2012.6 для поддержки научных исследований, проводимых ведущими научными школами РФ».

[2] Доклад публикуется в сокращенном варианте. Полная версия доступна на сайте конференции «Диалог».

(4) *Стало светло **и** тепло.*

Последний случай мы не будем рассматривать.

Коннекторы не являются отдельной частью речи. В этот класс объединяют некоторые союзы, некоторые частицы и некоторые наречия. Классический пример коннектора — *между тем*. Для простоты будем рассматривать данную единицу как одно слово, которое пишется с пробелом (в [МАС] оно квалифицируется как фразеологическое сочетание). Ср.

(5) *Мы уже давно сбились с дороги. Всем хотелось есть. Между тем начало смеркаться.*

Между тем связывает вводимое им высказывание (sentence) сразу с группой предшествующих высказываний. Что касается союза, то в примерах выше он связывает два предложения (т. е. clauses, в составе одного сложного), ср. (1)-(2), или два отдельных высказывания в тексте, ср. (3).

Разумеется, коннектор, подобно союзу, может связывать и просто два высказывания. Ср.

(6) *Мы уже давно шли по лесу без всякой дороги. Между тем начало темнеть.*

(7) *Мы второй день сидим без еды, между тем нам все время обещают помощь.*

Однако это частный случай. Главное свойство коннектора состоит в том, что он может связывать не только высказывания, но и т. н. «сверхфразовые единства».

Оказалось, что описание синтаксиса выбранных слов требует уточнения теории валентностей [Урысон 2008].

2. «Мягкие» и «жесткие» союзы

Некоторые союзы, подобно коннекторам, могут связывать целые группы предложений или высказываний. Это «мягкие» союзы. Другие союзы, напротив, способны соединять только два предложения (высказывания). Такие союзы мы называем «жесткими».

Мягкость/жесткость союза связана с еще одним, более слабым, свойством: способностью/неспособностью союза соединять два отдельных высказывания (т. е. sentences, а не clauses), ср. (3).

Логически возможны следующие комбинации.

(а) Союз способен связывать и два отдельных высказывания, и высказывание с группой высказываний. Это максимально мягкий союз.

(б) Союз способен связывать как два отдельных предложения (clauses), так и два отдельных высказывания (sentences), но не высказывание с группой высказываний. Это минимально мягкий союз.

(в) Союз способен связывать только два предложения (clauses), причем в составе сложного. Это жесткий союз.

По-видимому, многие сочинительные союзы — мягкие, а многие подчинительные союзы — жесткие.

Примеры.

(8) *Он сделал мне предложение, и я вышла за него замуж.*

(8а) *Он сделал мне предложение. И я вышла за него замуж.*

(9) *Он всегда помогал мне, был рядом в трудную минуту. Дети очень его любили. И я вышла за него замуж.*

В (8) союз *и* связывает два предложения, в (8а) — два высказывания, в (9) — группу предшествующих высказываний с данным (т. е. с тем, которое содержит союз *и*). Следовательно, союз *и* является максимально мягким.

Данные контексты с союзом *и* выражают причинно-следственное значение. Возьмем другие примеры, где союз *и* выражает т. н. «присоединительное» значение. Ср.

(10) *Снег тает, и день уже очень длинный.*

(10а) *Снег тает. И день уже очень длинный.*

(11) *Снег тает. По ночам коты орут. И день уже очень длинный.*

Ясно, что союз *и* в «присоединительном» значении тоже максимально мягкий.

Аналогичные примеры можно привести для союзов *но* и *а*.

Союз *а также*, подобно союзу *и*, является средством оформления множества, задаваемого перечислением. Ср. *Пришли Ваня, Петя, а также <и> Маша*. Элементами этого множества могут быть и ситуации. Ср. *Летят снаряды, а также свистят пули*; *Он сломал ногу, а также повредил руку*. При этом союз *а также* тяготеет к оформлению сложносочиненного предложения. Ср. неудачные примеры:

(12) ?*Петя сломал ногу. Он повредил руку. А также при падении у него выбило зубы.*

Союз *а также* является жестким — он может связывать только два простых предложения в составе сложного (т. е. sentences).

Основные причинные и уступительные союзы являются максимально мягкими. Ср.

(13) *Маша все делает для Пети. Она дает ему деньги в долг. Она помогала его маме в больнице. Она нашла для него хорошую работу. Потому что она до сих пор любит его.*

(14) *Мы долго и упорно делали эту работу. Сидели ночами. Тщательно проверяли расчеты. Хотя <несмотря на то что> эта работа нам всем уже давно надоела.*

Временные союзы *когда* и *пока* тоже могут связывать предложение с целой группой предшествующих высказываний. Ср.

(15) *Петя до одурения играл в компьютерные игры. Он болтал по телефону и не прикасался к учебникам. Когда отца не было дома.*

(16) *Петя до одурения играл в компьютерные игры. Он болтал по телефону и не прикасался к учебникам. Пока отца не было дома.*

Союзы *когда* и *пока* ведут себя здесь как максимально мягкие. Тем не менее, в интересующем нас отношении они отличаются от *хотя* и *несмотря на то что*. Действительно, сравним (15) и (16) со следующими примерами:

(17) *Петя до одурения играл в компьютерные игры, он болтал по телефону и не прикасался к учебникам, когда отца не было дома.*

(18) *Петя до одурения играл в компьютерные игры, он болтал по телефону и не прикасался к учебникам, пока отца не было дома.*

Ясно, что (15) и (16) выражают некоторый дополнительный модальный смысл. В первом приближении, этот смысл можно сформулировать так: ‘P, только когда <пока> Q’; кроме того, (15) и (16) акцентируют, что отец запрещал Пете делать все это. В примерах (17)–(18) этот смысл специально не выражен. Однако благодаря знанию действительности примеры (17)–(18) понимаются практически так же.

В тех контекстах, которые не допускают такого модального понимания, союз *когда* ведет себя как максимально жесткий: он может связывать только два предложения в составе сложного, но не два отдельных высказывания. Ср.

(19а) *Я закрыл книгу, когда стало совсем темно.*

(19б) ?? *Я закрыл книгу. Когда стало совсем темно.*

(Мы отвлекаемся от парцелляции и сегментирования предложения.)

Поведение союза *пока* сложнее: в некоторых контекстах, например в контексте императива, он может связывать и два отдельных высказывания, т. е. быть минимально мягким. Ср.

(20) *Уходи. Пока я тебя не побил.*

Жестким является союз *прежде чем*. Ср.

(21а) *Он остановился, прежде чем она приблизилась к нему.*

(21б) ?? *Он остановился. Прежде чем она приблизилась к нему.*

Последний пример возможен лишь с контрастным ударением на первом компоненте союза.

Жесткими являются также все те союзы, которые состоят из двух дистантных элементов; ср. *(не) то… (не) то, если…то, чем…тем* и т. п.

В класс коннекторов естественно только включать мягкие союзы.

3. Представление синтаксиса союзов и коннекторов

Союзы естественно считать предикатами с двумя семантическими актантами — P и Q, причем в нашем материале оба семантических актанта союза представляют собой ситуации. Ср. *Хотя был сильный мороз (Q), лыжные тренировки не отменялись (P)* [‘хотя Q, P’]; *Когда ты сдашь последний экзамен (Q), мы поедем на дачу (P)* [‘когда P, Q’]; *Шел дождь* (P)*, и было холодно* (Q) [‘P, и Q’]; *Я устала* (P)*, а тебе никакого дела нет* (Q) [‘P, а Q’].

Сколько у союза синтаксических валентностей?

Синтаксическая валентность лексемы — это, по определению, «способность предикатной лексемы синтаксически подчинять себе слово, группу слов или предложение, которые соответствуют ее семантическому актанту» [Апресян 2004: XXV]. Из этого определения следует, что синтаксический актант некоторой предикатной лексемы — это ее синтаксическое зависимое.

Условимся, что в рассматриваемых нами случаях как подчинительный, так и сочинительный союз синтаксически подчиняют себе оба связываемых предложения: *P ← когда → Q, P ← хотя → Q, P ← и → Q, P ← а → Q.*

В случае жесткого союза правила реализации его синтаксических валентностей принципиально не отличаются от, скажем, правил реализации валентностей глагола (с той разницей, что его синтаксические валентности насыщаются отдельными предложениями в составе сложного). У минимально мягкого союза синтаксический актант *P* может быть выражен либо предшествующим предложением в составе данного сложного, либо высказыванием, которое непосредственно предшествует данному. В этих случаях оба синтаксических актанта союза описываются в синтаксических терминах и могут быть полностью заданы синтаксическими правилами.

А как описать синтаксис коннектора или максимально мягкого союза?

Сравним подчинительный союз *хотя* и коннектор *тем не менее* [В. Апресян 1999, 2004; Урысон 2003].

(22а) *Был очень сильный мороз* (P). *Тем не менее лыжные тренировки не отменялись* (Q).

(22б) *Хотя был очень сильный мороз* (P), *лыжные тренировки не отменялись* (Q).

И союз *хотя*, и коннектор *тем не менее* описывают положение дел, когда имеют место две ситуации — P и Q, причем ситуация P препятствует существованию ситуации Q. Тем самым, слова *хотя* и *тем не менее* имеют по два семантических актанта. Актант P — это ситуация-препятствие, актант Q — «ключевая» ситуация. Однако союз *хотя* непосредственно вводит указание на ситуацию-препятствие (P) и располагается в соответствующем предложении. А *тем не менее* вводит указание на «ключевую» ситуацию (Q) и располагается в предложении, обозначающем данную ситуацию. Условимся, что слово *тем не менее* подчиняет себе вводимое им предложение. Легко видеть, что *тем не менее* и *хотя* находятся в отношении конверсии [Урысон 2003], [В. Апресян 1999].

Каждому семантическому актанту союза *хотя* соответствует синтаксический актант (один из них — главное предложение, а второй — придаточное), причем как оба семантических, так и оба синтаксических актанта союза *хотя* являются обязательными. Синтаксические актанты союза *хотя* описываются в синтаксических терминах и могут быть полностью заданы синтаксическими правилами. Союз *хотя* — это и семантически, и синтаксически двухместный предикат.

А сколько синтаксических валентностей (и синтаксических актантов) у слова *тем не менее*?

Теория валентностей (в частности, используемая в модели «СМЫСЛ↔ТЕКСТ») предназначена прежде всего для представления семантических и синтаксических актантов предиката внутри предложения [Мельчук 1974; Апресян 1974]. Рамками высказывания ограничивается и более широкая теория сфер действия лексических единиц [Богуславский 1996]. Между тем единица *тем не менее* принадлежит к классу коннекторов, а они выражают семантические связи между фрагментами текста, иногда достаточно большими. Функционирование таких единиц удается представить только выйдя за пределы высказывания.

Слово *тем не менее* располагается в начале или независимого высказывания или предложения (ср. *Тем не менее лыжные тренировки не отменялись* — *Лыжные тренировки тем не менее не отменялись*). Как мы условились, оно подчиняет себе это высказывание (предложение). Следовательно, слово *тем не менее* обладает синтаксической валентностью на «ключевую» ситуацию Q. Соответствующий синтаксический актант выражается предложением, в котором находится *тем не менее*.

Что касается ситуации-препятствия P, то ее обозначение может описываться весьма длинным фрагментом предтекста. Так, в текстах, откуда взяты следующие примеры, описание «препятствий» для Q занимает не одну страницу. Ср.

(23) *Тем не менее костюм от редакции я получил* (С. Довлатов).

(24) *Тем не менее друг к другу они чувствовали настоящую приязнь и разговаривали обо всем совершенно свободно* (Ю. Домбровский).

Эти примеры, между прочим, демонстрируют, почему коннектор естественно синтаксически относить к правому, а не к левому конъюнкту. (Можно было бы считать, что синтаксически коннектор вообще не относится ни к одному высказыванию текста. Однако подобный подход требует принципиального расширения синтаксической теории, без которого, как кажется, можно вполне обойтись.)

Из (23)–(24) ясно, что к выражению семантического актанта Р слова *тем не менее* не может быть применено определение синтаксической валентности и синтаксического актанта: синтаксический актант некоторой предикатной лексемы — это ее синтаксическое зависимое, а оно располагается в пределах того же высказывания, что и сам данный предикат.

Разумеется, ситуация-препятствие Р выражена в пределах того же высказывания, что и лексема *тем не менее*. Ср.

(25) *Подавленный всем случившимся, он тем не менее продолжает лекцию* [ситуация Р — ‘он подавлен всем случившимся’].

(26) *Мало что понимая по-немецки, Иван тем не менее догадывался, как издевается над всем виденным Георг Майзель* (А. Азольский) [ситуация Р — ‘Иван мало что понимал по-немецки’].

(27) *В этом городе, где уже столько лет вытравляется дух приключения, оно тем не менее живет, ползет по улицам, лепится к окнам* (В. Аксенов) [ситуация Р — ‘в этом городе уже столько лет вытравляется дух приключения’].

Однако ситуация-препятствие Р в приведенных примерах выражается самыми разными способами, причем невозможны никакие сколько-нибудь естественные соглашения, при которых соответствующие фрагменты предложения разумно было бы считать синтаксическими зависимыми слова *тем не менее*. Получается, что данная единица не имеет синтаксической валентности, соответствующей ее семантическому актанту Р.

Таким образом, союз *хотя* и семантически, и синтаксически двухвалентен, а слово *тем не менее*, имея два семантических актанта, должно быть признано синтаксически одновалентным. Между тем к обозначению его второго семантического актанта предъявляются вполне четкие формальные требования.

Во-первых, семантический актант «ситуация-препятствие» слова *тем не менее* выражается только в предтексте. Во-вторых, обозначение этого актанта должно содержать рему (основную или второстепенную), см. примеры выше. То есть какие-то формальные требования к выражению обсуждаемого

актанта есть, однако их слишком мало для автоматического анализа или синтеза.

В рамках модели «СМЫСЛ ↔ ТЕКСТ» (да и любой собственно синтаксической концепции) описание подобных ограничений на выражение семантического актанта не предусмотрено. Попытаемся применить к нашему случаю концепцию сфер действия лексических единиц [Богуславский 1996].

4. Обсуждение теоретического аппарата

Идея И. М. Богуславского состоит в том, что «разумно было бы называть синтаксическим актантом любое синтаксическое образование, значение которого соответствует семантическому актанту» [Богуславский 1996]. Но поскольку термин «синтаксический актант» имеет вполне устоявшееся, традиционное понимание, не распространяющееся на любое синтаксическое образование, то вместо него вводится новый, более широкий термин: синтаксическая сфера действия лексемы по ее семантическому актанту.

Казалось бы, приведенное определение идеально подходит для нашего случая: у слова *тем не менее* есть синтаксическая сфера действия по семантической валентности на ситуацию-препятствие P, а значит, и соответствующий синтаксический актант.

Однако во всех случаях, разобранных И. М. Богуславским, речь идет о тех синтаксических сферах действия, которые задаются четкими, вполне алгоритмизуемыми лингвистическими правилами. Это факт имеет принципиальный характер: если, заменяя понятие синтаксического актанта более широким и гибким понятием синтаксической сферы действия, мы откажемся от четких требований к поверхностному выражению семантического актанта, то нам придется отказаться и от идеи алгоритмизуемого обнаружения синтаксического представления высказывания.

Однако выражение семантического актанта P лексемы *тем не менее* не может быть задано подобным образом.

В рамках модели «СМЫСЛ ↔ ТЕКСТ» в таких случаях принято говорить, что данный семантический актант невыразим и лексема не имеет соответствующей синтаксической валентности. Ср. классический пример — глагол *промахнуться* [Мельчук 1974: 135; Апресян 1974: 148].

Глагол *промахнуться* описывает ситуацию, когда субъект стреляет (или бросает нечто) в цель, однако не попадает в нее. Однако семантический актант «мишень» не может оформляться как обычный синтаксический актант глагола. Причина в том, что в русском языке нет стандартного словосочетания ??*промахнуться по кому-л./чему-л. <*в кого-л./что-л.>*, хотя нормально словосочетание *стрелять в кого-л/что-л.* Данный «невыразимый» семантический актант часто выражается где-то в тексте. Ср. *Охотник выстрелил в волка, но промахнулся*. Однако выражение этого семантического актанта невозможно задать теми формальными правилами, которыми описывается реализация валентностей.

Семантический актант «ситуация-препятствие» коннектора *тем не менее* тоже не может быть полностью задан формальными синтаксическими правилами. Следовательно, этот актант тоже невыразимым.

Однако между *промахнуться* и *тем не менее* есть принципиальное различие.

5. Два типа «невыразимых» семантических актантов

Невыразимость объекта глагола *промахнуться* обусловлена тем, что данный его семантический актант фигурирует в пресуппозиции лексемы, а не в ассерции [Плунгян, Рахилина 2005]. Аналогичным образом обстоит дело с глаголами, обозначающими ликвидацию результата действия. «У глаголов типа *забинтовать (руку марлей), заворачивать (покупку в бумагу), завязывать (ящик веревкой)* имеется валентность средства-объекта, нереализуемая у их антонимов *разбинтовать (руку), разворачивать (покупку), развязывать (ящик)*, хотя в принципе она у них есть, потому что их невозможно истолковать, не упоминая некоего Х-а (средства-объекта), который А снимает с В. Аналогичную семантическую структуру имеют и другие антонимы этого класса, к числу которых можно отнести, помимо уже названных слов, *развьючить (осла), разгрузить (машину), разжать (уши), расстегнуть (пальто)* и многие другие (ср. исходные глаголы *навьючить (осла тюками), нагрузить (машину зерном), зажать (уши ладонями), застегнуть (пальто на все пуговицы)*).» [Апресян 1974: 147]. У всех этих слов тоже невыразим семантический актант, фигурирующий в пресуппозиции значения лексемы. Естественно, что такой подразумеваемый актант может вообще никак не называться в тексте, что не нарушает его правильности. Ср. *Не знаю, куда он метил, знаю только, что промахнулся*; *Чем он зарабатывает? — Вагоны разгружает* и т. п.

Что касается лексемы *тем не менее*, то выражение ее обсуждаемого актанта в тексте совершенно обязательно. Высказывание типа *Тем не менее костюм от редакции я получил* обязательно предполагает предтекст, содержащий описание ситуации-препятствия Р. Ситуация парадоксальна: семантический актант, с одной стороны, по определению «невыразим», а с другой стороны, его выражение в тексте совершенно обязательно.

Для представления рассмотренных фактов мы воспользуемся концепцией И. М. Богуславского, введя, однако, новые определения.

6. Типы синтаксических сфер действия лексемы по некоторому актанту

- Если синтаксическая сфера действия лексемы по некоторому семантическому актанту вполне описывается строгими лингвистическими правилами, будем говорить, что данная синтаксическая сфера лингвистически хорошо определена.

Хорошо определены синтаксические сферы действия, соответствующие «классическим» синтаксическим актантам, например глагольным дополнениям. Хорошо определенными являются и синтаксические сферы действия частиц, описанные в книге [Богуславский 1996].

- Если синтаксическая сфера действия лексемы по некоторому семантическому актанту не может быть описана строгими лингвистическими правилами, будем говорить, что данная синтаксическая сфера является лингвистически неопределенной.

Пример: синтаксическая сфера действия семантического актатнта «цель» глагола *промахнуться*. «Невыразимый» семантический актант некоторой лексемы — это такой актант, синтаксическая сфера действия по которому лингвистически не определена.

- Будем говорить, что синтаксическая сфера действия лексемы по данному семантическому актанту является частично лингвистически определенной, если выражение данного актанта описывается формальными правилами не полностью.

Частично определенной является синтаксическая сфера действия слова *тем не менее* по актанту «ситуация-препятствие».

Вернемся к синтаксическому различию между *хотя* и *тем не менее*. Оба слова имеют по два семантических актанта, и каждому актанту соответствует своя синтаксическая сфера действия. При этом у союза *хотя* обе синтаксические сферы действия являются лингвистически хорошо определенными (синтаксическая сфера действия по одному актанту — это главное предложение, а по другому актанту — придаточное). У частицы *тем не менее* лингвистически хорошо определена синтаксическая сфера действия только по одному актанту («ключевая ситуация»). Синтаксическая сфера действия по второму актанту («ситуация-препятствие») определена лишь частично.

У максимально мягких союзов, как и коннекторов типа *тем не менее, между прочим*, сфера действия по «левому» семантическому актанту тоже определена лишь частично.

Заключение

Для нахождения семантического актанта, которому соответствует частично определенная синтаксическая сфера действия, требуются не только собственно лингвистические сведения (например, о линейном расположении фрагментов текста), но и понятийный анализ текста, невозможный без привлечения обширных «энциклопедических знаний» о действительности. Возможно, подобная процедура и поддается какой-то алгоритмизации, но при

этом предполагается обработка не синтаксической или семантической структуры высказывания (текста), а некоторой понятийной сети.

Литература

1. *Апресян* 1974 — Ю. Д. Апресян. Лексическая семантика. М.
2. *Апресян* 2005 — Ю. Д. Апресян. Лингвистическая терминология Нового объяснительного словаря синонимов русского языка // Новый объяснительный словарь синонимов русского языка. 2-е изд., испр. и дополн. Под общим руководством акад. Ю. Д. Апресяна. М.
3. *Богуславский* 1996 — И. М. Богуславский. Сфера действия лексических единиц. М.
4. *В. Апресян* 1999 — В. Ю. Апресян. Уступительность в языке и слова со значением уступки // ВЯ. 1999. № 5.
5. *Крейдлин, Падучева* 1974а — Г. Е. Крейдлин, Е. В. Падучева. Значение и синтаксические свойства союза а // НТИ. Сер. 2. 1974, № 9.
6. *Крейдлин, Падучева* 1974б — Г. Е. Крейдлин, Е. В. Падучева. Взаимодействие ассоциативных связей и актуального членения в предложениях с союзом а // НТИ. Сер. 2. 1974, № 10.
7. *Мельчук* 1974 — И. М. Мельчук. Опыт теории построения моделей «Смысл ↔ Текст». М., 1974.
8. *Падучева* 1974 — Е. В. Падучева. Семантика синтаксиса.
9. *Плунгян, Рахилина* 2005 — V. Plungian, E. Rakhilina. Приземлиться и промахнуться: семантические механизмы синтаксических ограничений // Восток — Запад: Вторая международная конференция по модели «Смысл ↔ Текст». М. 2005.
10. *Теремова* 1986 — Р. М. Теремова. Семантика уступительности и ее выражение в современном русском языке. Л., 1986.
11. *Урысон* 2003 — Е. В. Урысон. Семантическая и валентная структура слов с уступительным значением // Русский язык в научном освещении. № 6.
12. *Урысон* 2008 — Е. В. Урысон. Синтаксические свойства союзов и теория валентностей // Динамические модели: Слово. Предложение. Текст. / Сборник статей в честь Е. В. Падучевой. М., 2008.

УПРАВЛЕНИЕ ИНОЯЗЫЧНЫХ СУЩЕСТВИТЕЛЬНЫХ — НАЗВАНИЙ ЛИЦ ПО ПРОФЕССИИ В СОВРЕМЕННОМ РУССКОМ ЯЗЫКЕ

Валиахметова А. Р. (episteltjes@gmail.com)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

В статье рассматриваются семантика и управление иноязычных слов — названий лиц по профессии в русском языке конца XIX — начала XXI века. Описаны три семантические группы: ‘сотрудник’ (*менеджер, агент, фрилансер, аутсорсер*), ‘художник’ (*дизайнер, кутюрье, стилист*) и ‘преподаватель’ (*репетитор, тьютор*). Они управляют либо беспредложным родительным падежом, либо предложно-падежной конструкцией *по чему-л*. На выбор варианта управления влияет ряд факторов: семантика слова, семантическая и словообразовательная аналогия. На данном этапе исследования выявлено, что существительные из групп ‘сотрудник’ и ‘преподаватель’ предпочитают управление конструкцией *по чему-л.*, что соответствует общей тенденции, действующей для слов — названий лиц с конца XIX в. Слова, относящиеся к группе ‘художник’, чаще управляют беспредложным родительным падежом под влиянием семантической аналогии с более ранним заимствованием *модельер*.

Ключевые слова: заимствования, актуальная лексика, вариативность, именное управление

Вопросы, связанные с употреблением иноязычных слов и их фиксацией в словарях, всегда вызывали исследовательский интерес. На рубеже XX–XXI веков они стали особенно актуальными в связи с интенсификацией процесса заимствования в русском языке.

В центре внимания лингвистов, изучающих иноязычные слова, обычно находятся социолингвистические вопросы и особенности освоения заимствований: фонетического, орфографического, словообразовательного, грамматического и семантического. Синтаксические свойства заимствований обычно остаются за пределами рассмотрения исследователей. Тем не менее, очевидно, что сочетаемость иноязычных слов с другими словами вызывает у говорящих не меньше трудностей, чем, например, их правописание и произношение. На это указывает Л. П. Крысин: «Зона иллюстраций — в виде словосочетаний, речений — особенно актуальна для лексикографического представления неологизмов, поскольку у носителей языка трудности могут возникать как раз при употреблении иноязычного неологизма в тексте, при его соединении с другими словами в составе предложения» [Крысин 2008: 97].

Данная статья посвящена управлению иноязычных слов, вошедших в русский язык или актуализировавшихся в конце XX — начале XXI века, а именно существительных — названий лиц по профессии. Многие исследователи отмечают, что в последние годы список этих названий интенсивно пополняется иноязычными заимствованиями, а М. А. Кронгауз даже относит некоторые из них к ключевым словам эпохи [Кронгауз 2007]. При этом управление большинства этих слов в современных словарях русского языка не отражено; для некоторых синтаксическая информация приводится в иллюстративных примерах (например, *агент по снабжению* (БТС[1]), *менеджер по рекламе* (НСИС)). В статье будет рассмотрено управление иноязычных слов *менеджер, агент*, *дизайнер* и *репетитор*, отмеченных в словарях актуальной лексики (ЯИ, АЛ), а также неологизмов *аутсорсер, фрилансер, кутюрье, стилист* и *тьютор*. Семантически они делятся на три группы: ‘сотрудник’ (*менеджер, агент, фрилансер, аутсорсер*), ‘художник’ (*дизайнер, кутюрье, стилист*) и ‘преподаватель’ (*репетитор, тьютор*).

1. «Сотрудники» (*менеджер, агент, фрилансер, аутсорсер*)

Прежде чем обратиться к управлению этой группы слов, рассмотрим сначала их значение и функционирование в современном русском языке.

Все эти существительные встречаются на сайтах, посвящённых поиску вакансий и персонала[2], в качестве наименований профессий. Следует отметить, что слова *агент* и *менеджер* используются как контекстные синонимы. Например, в заголовке вакансии значится «Ведущий менеджер по продажам», а в самом тексте — «ведущий агент по продаже»[3].

Семантика существительного *менеджер,* как уже отмечали многие лингвисты, стала весьма размытой: если раньше оно означало специалиста по управлению производством (ОШ, БТС, ИСИС), то теперь это любой сотрудник компании [Кронгауз 2007: 57]. Этой размытостью значения обусловлена сочетаемость слова *менеджер* с самым широким кругом существительных. Это отмечают как лингвисты: *менеджер по клинингу* (МС, [Кронгауз 2007: 57]), так и обычные носители языка: *Можно и уборщицу назвать* ***менеджером по уборке*** *помещений — сути дела это не изменит* (Форум, 2006). У многих словосочетаний со словом *менеджер* имеются эквиваленты в кодифицированном литературном языке (*менеджер по закупкам — снабженец, менеджер по персоналу — кадровик*) и в разговорной речи (*менеджер по рекламе — рекламщик, менеджер по продажам — продажник*). То же можно сказать и об *агенте*: *агент по доставке — курьер, агент по аренде / агент по недвижимости — риелтор*.

---

[1] Здесь и далее расшифровку аббревиатур см. в приложении «Словари».

[2] http://hh.ru, http://rabota.ru, http://job.ru, http://career.ru.

[3] Вакансия на сайте http://www.superjob.ru/, 2012.

Слова *фрилансер* (от англ. *freelancer*, букв. «свободный художник») и *аутсорсер* (от англ. *outsourcer*, букв. «внешний источник») примыкают к словам *менеджер* и *агент*, т. к. они тоже обозначают сотрудников компании. Однако их семантика несколько у́же: *фрилансер* и *аутсорсер* — это внештатные сотрудники, работающие, как правило, вне офиса. При этом сфера деятельности фрилансеров обычно ограничена «журналистикой.., компьютерным программированием и дизайном.., переводом, разного рода консультационной деятельностью, частной фото- и видеосъемкой» (Википедия). Аутсорсеров же нанимают для «ведения бухгалтерского учёта, обеспечения функционирования офиса, переводческих услуг, транспортных услуг, поддержки работы компьютерной сети и информационной инфраструктуры, рекламных услуг, обеспечения безопасности» (там же). В связи с этим лексическая сочетаемость этих существительных не столь широка, как у слов *менеджер* и *агент*.

Перейдём теперь к управлению слов данной группы.

1.1. В толковых словарях при существительном *менеджер* приводятся примеры с управлением предложно-падежной конструкцией *по чему-л.*: *м. по рекламе, по подбору персонала* (НСИС); *м. по рекламе, по работе с ценными бумагами* (ЯИ). Как отмечают лингвисты, это управление уже стало устойчивым: *менеджер по дилерским операциям* [Подчасова 1994], *менеджер по продажам, по работе с клиентами* [Кронгауз 2007: 57].

В узусе встречаются словосочетания с разным лексическим наполнением: *менеджер по торжествам, по тиражу, по алкоголю, по качеству, по рекламациям* и т. п., например:

(1) ***Менеджерам по торжествам*** *по силам разработать сценарий любого мероприятия*[4].

(2) *По словам категорийного* ***менеджера по алкоголю*** *сети «Гигант» Алексея Ильюшенка, на прошлой неделе действительно был перерыв в поставках продукции «ВИНАПа»*[5].

Встречается также управление беспредложным родительным падежом: *м. продаж, м. закупок*[6].

Этот вариант употребляется гораздо реже (ср., например, по поиску в Яндекс.Новостях[7]: 17 663 вхождения словосочетания *менеджер по продажам*; 421 — *менеджер продаж*).

---

[4] Новостной сайт, 2006.

[5] Национальный корпус русского языка (далее — НКРЯ), 2004.

[6] Вакансии на сайте http://hh.ru, 2011, 2012.

[7] Поиск по сегменту «Яндекс.Новости» осуществлялся в конце марта 2012 г. Анализ слова *менеджер* в качестве управляющего слова при помощи Яндекса, как нам представляется, не является репрезентативным, т. к. выдает слишком большое количество

1.2. Для существительного *агент* также закрепилось управление конструкцией *по чему-л.*, о чём свидетельствуют как иллюстративные примеры в словарях русского языка (*агент по снабжению* (БТС), *агент по недвижимости* (ИСИС)), так и заголовки статей в словарях, посвящённых сфере бизнеса и финансов, например:

(3) ***Агент по продаже*** *— агент по найму, который отслеживает распространение товаров фирмы заказчика в определенном регионе*[8].

Сочетание *агент по чему-л.* встречается также на сайтах, посвящённых поиску вакансий и персонала: *а. по аренде, по недвижимости, по продаже, по туризму* и мн. др.

Можно также отметить периферийный вариант управления: *агент чего-л.*: *а. продаж*[9], *а. доставки*[10]. Он значительно менее распространен (ср., например, по поиску в Яндекс.Новостях: 737 вхождений словосочетания *агент по продажам*; 42 — *агент продаж*).

1.3. Существительное *фрилансер* не сопровождается в словарях иллюстративными примерами. Более того, согласно толкованиям, это слово не может управлять конструкцией с зависимым словом, обозначающим сферу деятельности, т. к. в его значении она уже содержится: фрилансер — это либо «журналист, самостоятельно выбирающий темы своих будущих публикаций, не связанный с определённой редакцией и пишущий на свой страх и риск без гарантии опубликования», либо «внештатный корреспондент», либо «творческий работник (писатель, переводчик и др.), работающий без договора» (НСИС).

Однако на сайтах, посвященных поиску вакансий и персонала, *фрилансер* употребляется в более общем значении ‘внештатный сотрудник’ и управляет зависимым словом, обозначающим сферу деятельности сотрудника. При этом

вхождений (например, **менеджер /+1 «по продажам»** — 40 млн страниц). Т. к. реальная выдача ограничена тысячью единиц, объявленное Яндексом число найденных документов может быть на порядки больше или, наоборот, меньше (из-за дублирования некоторых страниц).

То же касается и некоторых других рассматриваемых далее слов, например *дизайнер* и *стилист*. Так, при запросе со словом *стилист* выдача Яндекса явно носит рекламный характер: выдаётся множество ссылок на салоны красоты, в описаниях услуг которых очень распространена именно конструкция *по чему* (ср. [Левонтина 2010: 206]: *программа по телу, уход по телу, уход по лицу*). Следовательно, эта статистика не может быть учтена, т. к. искажает языковую реальность.

Поиск по сегменту «Яндекс.Новости» во-первых, дает вполне обозримые результаты, во-вторых, исключает явно ненормативные контексты, удалить которые при выдаче на Яндексе более 1000 контекстов невозможно.

[8] Словарь бизнес-терминов, 2001 // Словари на Академике, http://dic.academic.ru/.

[9] Вакансия на сайте http://career.ru, 2012.

[10] Вакансия на сайте http://www.job.ru, 2011.

используется либо конструкция *по чему-л.*: *ф. по верстке*[11], *ф. по дизайну интерьеров*[12]; либо беспредложный родительный падеж: *ф. верстки*[13], *ф. дизайна интерьеров*[14].

Предложно-падежная конструкция используется с более широким кругом слов: *ф. по рекламе*[15], *ф. по монтажу видеороликов*[16], *ф. по написанию статей*[17] и мн. др.

Как показывает статистика[18], в случае если возможны обе конструкции, более употребителен вариант *по чему-л.*:

по верстке	*верстки*	*по дизайну*	*дизайна*	*по переводу*	*перевода*
13	1	47	8	13	3

1.4. Существительное *аутсорсер* встречается лишь в одном словаре русского языка (СНТСРЯ). Оно толкуется через слово *аутсорсинг*, которое зафиксировано также в АЛ и НСИС.

Как уже было отмечено, сочетаемость слова *аутсорсер* у́же, чем у слов *менеджер* и *агент*, что обусловлено его семантикой. Встречаются в основном словосочетания, где зависимые слова относятся к финансовой сфере, при этом используются беспредложная и предложная конструкции: *а. продаж*[19], *а. бухгалтерии*[20]; *а. по продажам*[21], *а. по бухгалтерии*[22].

По статистике Яндекса затруднительно делать выводы о тенденциях в синтаксическом поведении этого неологизма:

---

[11] Резюме на сайте http://edu.jobsmarket.ru.

[12] Резюме на сайте http://www.free-lance.ru/, 2011.

[13] Резюме на сайте http://www.free-lance.ru/, 2011.

[14] Блоги, 2010.

[15] Форум, 2007.

[16] Форум, 2011.

[17] Резюме на сайте http://www.telejob.ru, 2011.

[18] Осуществлялся поиск по Яндексу; запросы вида **фрилансер /+1 «верстки»** вводились в конце марта 2012 г. Следует отметить, что количество страниц, выдаваемое Яндексом в левом верхнему углу, не соответствовало реальному. Было произведено отсеивание неадекватных запросу страниц вручную.

[19] Интернет-справочник компаний, http://www.vl.ru/callsales.

[20] Вакансия на сайте http://moikrug.ru/, 2010.

[21] «Справочник экономиста» №12 (90) декабрь 2010.

[22] Форум, 2011.

[23] Осуществлялся поиск словосочетаний *аутсорсер по бухгалтерскому учету*, *по бухгалтерии* и *по бухучету*. То же для управления беспредложным родительным падежом.

по продажам	*продаж*	*по бухгалтерскому учету*[24]	*бухгалтерского учета*	*по маркетингу*	*маркетинга*
3	34	14	11	0	5

Таким образом, устоявшимся в узусе управлением для слов *менеджер, агент* и *фрилансер* является конструкция *по чему*. При этом для слов *менеджер* и *агент* это управление фиксируется в словарях. Это соответствует общей тенденции, действующей для слов — названий лиц с конца XIX в. [Белошапкова 1964: 194]. Управление неологизма *аутсорсер,* означающего также сотрудника компании, колеблется между предложной конструкцией *по чему-л.* и беспредложным родительным падежом.

2. «Художники» (*дизайнер, кутюрье, стилист*)

Слово *дизайнер* вошло в русский язык предположительно во второй половине XX в.: в СУш его нет, а первое вхождение в НКРЯ датируется 1960 г. В словарях оно толкуется как ‘художник-конструктор, специалист по дизайну’ (МАС, ОШ, АЛ, НСИС, ИСИС).

Существительное *кутюрье* вошло в русский язык несколько позже: в МАСе его нет, первое вхождение слова в НКРЯ датируется 1970–1980 гг. Оно фиксируется в ряде современных толковых словарей, где толкуется как ‘модельер, имеющий собственное дело и высокого класса по проектированию одежды’ (НСИС, ИС, ННЛ), ‘модельер одежды, отвечающей требованиям современной моды’ (ИСИС).

Слово *стилист* является вторичным заимствованием: в его первичном значении ‘человек, владеющий искусством литературного стиля, пишущий хорошим стилем’ (СУш, СШ) оно заимствовано из греческого (по-видимому, через посредство французского), а в новом ‘художник-дизайнер, специалист, разрабатывающий стильные образцы одежды, аксессуаров, причёсок, макияжа’ (СШ) — из английского. Первое вхождение в НКРЯ в этом значении датируется 1996 г. Новое значение зафиксировано лишь в словарях начала XXI в. (ЯИ, АЛ, СШ, НСИС, ИСИС), а в некоторых (например, БТС) по-прежнему указано только первичное.

Рассмотрим теперь управление этих слов.

2.1. *Дизайнер* обычно управляет родительным падежом: *д. одежды*, *д. интерьера, д. сайтов*[24]. Встречается также и управление конструкцией *по чему-л.*: *д. по интерьеру, д. по шторам, д. по аксессуарам*[25].

---

[24] Вакансии на сайте http://www.job.ru, 2011.

[25] Вакансии на сайте http://www.hh.ru, 2012.

По статистике Яндекс.Новостей[26], преобладает беспредложное управление:

по обуви	*обуви*	*по одежде*	*одежды*	*по интерьеру*	*интерьера*	*по автомобилям*	*автомобилей*
2	661	13	5420	239	1151	1	184

Управление родительным падежом поддерживается аналогией со словом *модельер* — более ранним заимствованием, не столь актуальным в последние десятилетия. Согласно иллюстративным примерам в словарях, это существительное тоже управляет родительным падежом: *м. женского платья* (СШ), *м. дамского платья* (БТС)[27].

Возможно, помимо семантической аналогии, действует и словообразовательная: так, существительное *дизайн* употребляется с родительным падежом: например, *дизайн автомобиля* (БТС).

2.2. *Кутюрье* управляет родительным падежом:

(4) *Под стать знаменитым* ***кутюрье одежды*** *я ежегодно создаю новые коллекции обоев*[28].

(5) *Почему в России исчезли* ***кутюрье обуви*** *и есть ли надежда на их возрождение?*[29]

(6) *В тех или иных вариациях замки и пряжки включили в отделку все* ***кутюрье аксессуаров***[30].

Как видно из примеров, происходит расширение значения данного неологизма: согласно толкованию в словарях, кутюрье создает одежду, а примеры из узуса свидетельствуют о том, что он может создавать также обувь, аксессуары и даже интерьер:

(7) *Современные* ***кутюрье интерьера*** *сделали нынешний тренд более привлекательным и декоративным*[31].

---

[26] Поиск по сегменту «Яндекс.Новости» осуществлялся в конце марта 2012 г. См. также Прим. 7.

[27] В узусе встречаются также немногочисленные примеры управления сочетанием *по чему-л.*: *м. по одежде* (Вакансия на сайте http://job.webrostov.ru,2011), *м. по обуви* (Вакансия на сайте http://www.superjob.ru/, 2009)

[28] Новостной сайт, 2010.

[29] Новостной сайт, 2005.

[30] Женский интернет-журнал, 2005.

[31] Журнал «Городское обозрение», 2002–2012.

По статистике Яндекса[32], в словарном значении *кутюрье* управляет только беспредложным родительным падежом:

по одежде	*одежды*	*по детской одежде*	*детской одежды*	*поженской одежде*	*женской одежды*
0	23	0	7	0	22

В расширенном значении (например, со словами *прическа, аксессуары*) рассматриваемый неологизм может также управлять предложно-падежной конструкцией *по чему-л.*:

(8) *Вы просто великий* ***кутюрье по аксессуарам!***[33]

(9) *Чем отличается работа* ***кутюрье по прическам*** *от работы парикмахера или стилиста по волосам?*[34].

Интересно отметить, что, по статистике Яндекса, *кутюрье по прическам* встречается значительно чаще, чем *кутюрье причесок*:

по прическам	*причесок*
158	5

2.3. У существительного *стилист* в узусе конкурируют два варианта управления: беспредложным родительным: *с. волос, с. макияжа, с. маникюра, с. одежды*[35] и сочетанием *по чему-л.*: *с. по обуви*[36], *с. по прическам*[37], *с. по одежде*[38].

По статистике Яндекс.Новостей, преобладающим для слова *стилист* является управление конструкцией *по чему-л.*:

по прическам	*причесок*	*по макияжу*	*макияжа*	*по маникюру*	*маникюра*	*по одежде*	*одежды*
65	19	3	0	2	0	21	5

[32] Вхождений в Яндексе слова *кутюрье* с управляемыми словами не так много, поэтому поиск велся по всему поисковику, а не по сегменту «Яндекс.Новости» (хотя тоже потребовалась вычитка примеров: например, по запросу ***кутюрье /+1 «одежды»*** выдано 323 ответа; адекватных примеров оказалось 23).

[33] Блоги, 2010.

[34] Женский интернет-журнал, 2010.

[35] Женский интернет-журнал, 2011.

[36] Домовой, 1993. Пример из ЯИ.

[37] Женский интернет-журнал, 2011.

[38] Новостной сайт, 2012.

Итак, для слов из группы ‘художник’ характерно управление беспредложным родительным падежом и предложно-падежной конструкцией *по чему-л*. Эти варианты дублетны. У слов *дизайнер* и *кутюрье* преобладает управление беспредложным родительным, а конструкция *по чему-л*. является периферийной, что поддерживается аналогией со словом *модельер*. Существительное *стилист* чаще управляет сочетанием *по чему-л*.

3. «Преподаватели» (*репетитор, тьютор*)

Эти существительные имеют разное время появления в русском языке и разные языки-источники.

Так, *репетитор* заимствован из латыни (repetitor — тот, кто повторяет) и отмечен ещё в словаре Даля («помощник учителя, для повторенья уроков с учениками»). Первое вхождение, отмеченное в НКРЯ, — 1848 г.

В первой половине XX в. репетиторами называют домашних учителей, которые помогают неуспевающим ученикам: об этом свидетельствует толкование в СУш и употребления в узусе, например:

(10) *А мне,* ***балбесу, лентяю и грубияну, не могшему одолеть первых начал алгебры****, разве не нанимала она репетиторов, или, как она сама называла,* ***«погонялок»***[39].

Однако во второй половине XX в. учениками репетиторов становятся не отстающие, а любые школьники, чаще всего — готовящиеся к поступлению в вуз (АЛ), ср.:

(11) *Полгода репетиторы ходили. Все в один голос:* ***мальчик очень одарённый***[40].

Что касается существительного *тьютор*, то оно относится к неологизмам, заимствованным из английского языка. Оно зафиксировано лишь в современных словарях иноязычных слов, при этом квалифицируется как экзотизм (толкуется через топоним): «В системе школьного и университетского образования Великобритании: наставник, воспитатель; куратор студенческой группы» (ИСИС). В НСИС, помимо этого значения, указано также неэкзотическое — «преподаватель-консультант в системе дистантного обучения». В современном узусе *тьютор* употребляется также в значениях ‘научный руководитель’, ‘частный преподаватель’, ‘преподаватель, проверяющий тесты и другие письменные работы (в отличие от преподавателя-лектора) в системе заочного обучения’, ‘психолог, ведущий занятие-практикум в группе’, ‘методист’ [Маринова 2003].

Обратимся к управлению данных существительных.

---

[39] НКРЯ, 1932.

[40] НКРЯ, 2000.

3.1. В XIX в. репетитор употребляется в основном без управления. Поиск по НКРЯ дал лишь один пример:

(12) *Саша её **репетитор по занятиям** медициною*[41].

В XX в. наблюдается конкуренция двух вариантов: управления беспредложным родительным и конструкцией *по чему-л.*:

(13) *Посмотреть на него сзади — просто почтовый чиновник, спереди — из просителей, а не то **репетитор уроков***[42].

(14) *А наутро пришел **репетитор по языку**, и бабушка ему отказала*[43].

(15) *Многие просто упускают время, вследствие чего домашний **репетитор по физике** вынужден работать в режиме цейтнота*[44].

(16) *Хороший **репетитор математики** может научить Вашего ребенка не только логически мыслить, но и решать сложнейшие задачи*[45].

Носители отмечают эту вариативность и рефлексируют о ней в поисках «правильного» варианта:

(17) *Лучше сказать «**репетитор по английскому языку**» и «преподаватель английского языка», а не наоборот: «**репетитор английского языка**» или «преподаватель по английскому языку». Вы же не говорите «частный учитель по английскому языку». Скорее скажете «частный учитель английского языка»*[46].

Вопросы на справочном интернет-портале «Грамота.ру»: «Скажите, пожалуйста, как правильно: репетитор по русскому языку или репетитор русского языка? Если возможно, дайте ссылку на словарь или справочник»[47]; «Здравствуйте! Как правильно: “репетитор английского языка” или “реп. по англ. языку”?»[48]. «Грамота.ру» рекомендует предложно-падежное управление.

---

41 НКРЯ, 1863.

42 НКРЯ, 1923.

43 НКРЯ, 1959–1964.

44 Сайт «Ваш репетитор».

45 Сайт «Репетит-Центр».

46 Блоги, 2010.

47 Вопрос № 204877.

48 Вопрос № 249370.

3.2. Существительное *тьютор* встречается в узусе с теми же двумя вариантами управления: *т. математики и информатики*[49], *т. английского языка*[50]; *т. по английскому языку*[51], *т. по физике*[52].

Конкуренция этих двух вариантов прослеживается ещё в XIX в. при словах, называющих понятия, связанные со школой и образованием. Родительный беспредложный устойчиво употреблялся на протяжении всего XIX в.: *лекция русской истории, лекция географии; учитель этих предметов* [Белошапкова 1964: 195]. Кроме того, встречалась конструкция *в чём-л.*: *учитель в стихотворном языке* [Белошапкова 1964: 212], однако к концу XIX в. она затухает и распространяется сочетание *по чему-л*. Конкуренция беспредложного и предложного управления сохраняется и в современном русском языке, что отражено словарно: *Учитель математики, физики. Учитель по русскому языку, по природоведению* (БТС). Судя по словарному закреплению управления *репетитор по чему-л.* (БТС), а также оценке носителей этого варианта как правильного, можно делать выводы о дальнейшем вытеснении беспредложного родительного при словах данной группы. Это подтверждается такими словами, как *тренер* и *инструктор*, при которых возможно только сочетание *по чему-л.*: *инструктор по подготовке лётных кадров* (НСИС), *тренер по самбо* (ОШ), а также статистикой Яндекс.Новостей (для слова *репетитор*) и Яндекса (для слова *тьютор*):

Репетитор

по математике	*математики*	*по русскому языку*	*русского языка*	*по физике*	*физики*
293	9	91	21	48	3

Тьютор

по математике	*математики*	*по русскому языку*	*русского языка*	*по физике*	*физики*
73	9	90	4	31	3

А. М. Пешковский писал: «Нормой признаётся то, что было, и *отчасти то, что есть*, но отнюдь не то, что будет» [Пешковский 1959: 55, курсив мой — А. В.]. Действительно, норма как результат кодификации консервативна: она имеет ретроспективный характер и часто отстаёт от современного узуса.

[49] Резюме на сайте http://www.superjob.ru, 2008.

[50] Сайт частных объявлений.

[51] Сайт репетиторов.

[52] Форум, 2011.

Задачей данной работы было описание конкуренции существующих вариантов управления определённой семантической группы существительных — языкового материала, не отражённого консервативной нормой. Наблюдения над узусом позволяют говорить о двух вариантах управления этих слов: беспредложным родительным падежом и предложно-падежной конструкцией *по чему-л*. При этом на данном этапе изучения «сотрудники» и «преподаватели», следуя тенденции, начавшейся в XIX в., склоняются ко второму варианту (*агент по продаже, менеджер по рекламе, фрилансер по вёрстке, репетитор по русскому языку, тьютор по математике*), а «художники», подчиняясь аналогии со сходным по значению словом *модельер* — к первому (*дизайнер обуви, кутюрье одежды*).

Однако некоторые из рассмотренных слов не вписываются в данную тенденцию: так, у неологизма *аутсорсер* конкурируют оба варианта управления, и на данном этапе освоения слова выявить преобладающий не удается. Существительное *стилист*, в отличие от других «художников», управляет преимущественно сочетанием *по чему-л*.

Представляется, что на выбор варианта управления, помимо семантической и словообразовательной аналогии, влияют также тенденции семантического развития слова и связанное с этим расширение сочетаемости.

Так, рассмотрим группу «художники». В исходном значении существительное ***кутюрье*** управляет только словами, обозначающими одежду; при этом используется беспредложное управление родительным падежом. При употреблении слова в несловарном, более широком значении и, следовательно, с более широкой сочетаемостью, оно может управлять также конструкцией *по чему-л*. Преимущественное употребление существительного ***стилист*** с предложно-падежной конструкцией *по чему-л*. обусловлено также его широкой сочетаемостью (*прическа, макияж, одежда, аксессуары, интерьер* и мн. др.). Слово ***дизайнер*** тоже управляет разными по семантическим характеристикам лексемами (*одежда, интерьер, обувь, автомобили, шторы* и мн. др.), однако факторами, сдерживающими экспансию предлога *по*, являются семантическая (*модельер платья*) и словообразовательная (*дизайн автомобиля*) аналогии.

Обратимся теперь к группе «сотрудники». Как уже было отмечено, значение слов *менеджер* и *агент* весьма размыто, поэтому они управляют конструкцией *по чему-л*. Сравним неологизмы *аутсорсер* и *фрилансер*. Как видно из примеров, сочетаемость слова *фрилансер* шире, т. к. оно употребляется в общем значении ‘внештатный сотрудник’. Что касается сочетаемости существительного *аутсорсер*, то она ограничена словами, относящимися к экономической сфере (*бухгалтерский учет, продажи, маркетинг*). Вероятно, с этим связана меньшая экспансия предлога *по* в управлении этого неологизма.

Что касается группы «преподаватели», то, с одной стороны, их узкая сочетаемость (со словами, обозначающими учебные предметы) требует беспредложного управления, с другой стороны — аналогия со словом *учитель*, при котором ещё с конца XIX в. всё более употребительной становилась конструкция *по чему-л*., обусловливает вытеснение беспредложного родительного.

Таким образом, при изучении синтаксического поведения иноязычных неологизмов необходимо учитывать действие различных факторов: семантической аналогии, словообразовательной аналогии, а также семантики неологизма.

Литература

1. *Белошапкова В. А.* Изменения в субстантивных словосочетаниях // Изменения в системе словосочетаний в русском литературном языке XIX века / Под ред. В. В. Виноградова и Н. Ю. Шведовой. — М.: Наука, 1964. С. 128–220.
2. *Кронгауз М. А.* Русский язык на грани нервного срыва. — М.: Языки славянских культур, 2007. 232 с.
3. *Крысин Л. П.* Проблемы представления новых иноязычных заимствований в нормативных словарях // Крысин Л. П. Слово в современных текстах и словарях: Очерки о русской лексике и лексикографии. — М.: Знак, 2008. С. 83–97.
4. *Левонтина И. Б.* Русский со словарём. М.: Азбуковник, 2010. 335 с.
5. *Маринова Е. В.* Семантическая деривация в сфере иноязычной неологики // Проблемы языковой картины мира на современном этапе. Сб. статей. Вып. 2. — Н. Новгород: Изд-во НГПУ, 2003. С. 61–63.
6. *Пешковский А. М.* Объективная и нормативная точка зрения на язык // Избранные труды. — М.: Учпедгиз, 1959. С. 50–61.
7. *Подчасова С. В.* Брокеры, дилеры и другие посредники. — Русская речь, №№ 5, 6, 1994.
8. *Северская О. И.* Говорим по-русски с Ольгой Северской. — М.: Слово, 2004. 256 с.

Словари

1. *АЛ* — Толковый словарь русского языка начала XXI века. Актуальная лексика / Под ред. Г. Н. Скляревской. — М.: Эксмо, 2007. — 1136 с.
2. *БТС* — Большой толковый словарь русского языка / Под ред. С. А. Кузнецова. Электронный ресурс. Режим доступа: http://www.gramota.ru/slovari/info/bts/, свободный. Загл. с экрана. Данные соответствуют 05.02.2012. Словарь опубликован в авторской редакции 2009 года.
3. *ИС* — Окунцова Е. А. Новейший словарь иностранных слов. — М.: Айрис-пресс, 2009. — 512 с.
4. *ИСИС* — Крысин Л. П. Иллюстрированный словарь иностранных слов. — М.: Эксмо, 2011. 864 с.
5. *МС* — Новиков В. И. Словарь модных слов. — М.: Аст-пресс, 2011. 256 с.
6. *ННЛ* — Григоренко О. В. Новые наименования лиц в современном русском языке. Словарные материалы. — Воронеж: Научная книга, 2009. — 517 с.

7. *НСИС* — Захаренко Е. Н., Комарова Л. Н., Нечаева И. В. Новый словарь иностранных слов. — М.: Азбуковник, 2008. 1040 с.
8. *ОШ* — Ожегов С. И., Шведова Н. Ю. Толковый словарь русского языка. — М.: 1993.
9. *СНТСРЯ* — Шагалова Е. Н. Самый новейший толковый словарь русского языка XXI века. — М.: АСТ, 2011. — 413 с.
10. *СУш* — Толковый словарь русского языка: В 4 т. / Под ред. Д. Н. Ушакова. — М.: Гос. ин-т «Сов. энцикл.»; ОГИЗ; Гос. изд-во иностр. и нац. слов, 1935–1940.
11. *СШ* — Толковый словарь русского языка с включением сведений о происхождении слов / Под ред. Н. Ю. Шведовой. — М.: Азбуковник, 2007. — 1175 с.
12. *ТСЖВЯ* — Даль В. И. Толковый словарь живого великорусского языка: В 4 т. — М.: Русский язык, 1999.
13. *ЯИ* — Толковый словарь русского языка конца XX века. Языковые изменения / Под ред. Г. Н. Скляревской. — СПб.: Фолио-Пресс, 2000. — 700 с.

ЭКСПЕРИМЕНТАЛЬНЫЙ ПОДХОД К ОПРЕДЕЛЕНИЮ КОММУНИКАТИВНОЙ СТРУКТУРЫ ТЕКСТА

Ягунова Е. В. (iagounova_elena@mail.ru)

СПбГУ, Санкт-Петербург, Россия

В докладе рассматриваются возможности экспериментального подхода к исследованию коммуникативного членения высказываний и – шире – текстов. Анализируются тексты в звучащей vs. письменной форме; делового vs. художественного функционального стиля. Используется три разные методики членения: группой экспертов на материале звучащего текста, группой информантов и в ходе вычислительного эксперимента на материале письменного текста. Обобщение результатов всех видов экспериментов предоставляет возможность выделить интегральные коммуникативные структурные составляющие (КСС) текста. Тематические КСС, выделенные на основании всех трех экспериментов, соотносятся с основными ключевыми словами этих текстов, они же маркируют композиционные фрагменты текста и их темы. Для делового текста (и во многих случаях для художественного текста) КСС не противоречат правилам синтаксического структурирования. Особые КСС (в случае художественного нарратива) могут относиться скорее к структурированию ситуации, а не синтаксической структуры, что может подчеркиваться как интонационные кульминационные моменты ситуации.[1]

Ключевые слова: анализ текста, эксперимент, структура текста, тема, рема, коммуникативная структура

1. Введение. Терминологические проблемы. Постановка задачи

Самая глубинная (и в то же время самая грубая) структура, которую можно приписать любому предикативному смысловому образованию, — это тема-рематическая структура как сопряжение основных коммуникативных компонентов высказывания (равно и текста) (Хоккет 1970: 70).

В литературе многократно воспроизводилось определение темы (топика) в противопоставлении реме (фокусу, комментарию и т. п.), по сути повторяющее дефиниции субъекта и предиката в классической аристотелевской логике (ср. Касевич 2009). Некоторая размытость определения темы и ремы порождает многочисленные попытки их реинтерпретации.

[1] Доклад публикуется в сокращенном варианте. Полная версия доступна на сайте конференции «Диалог».

Во-первых, «топик» и «тема» некоторыми авторами различаются в зависимости от того, в описании структур какого уровня (напр., высказывания и/или текста) они участвуют (ср. Касевич 2009). Вслед за В. Б. Касевичем я использую эти термины как синонимы. Важно, что основной объект — текст, а коммуникативная и смысловая структуры всегда сложным образом взаимодействуют в процессах функционирования (порождения и восприятия текста). Очевидно, в рамках таких динамических процессов происходит взаимодействие коммуникативной и смысловой структур[2].

Типы взаимодействия в существенной степени зависят от функционального стиля текста, а далее от типа текста (напр., соотношения ролей функций передачи информации и воздействия на адресата, соотношения выраженной в тексте (эксплицитной) и явно невыраженной (имплицитной или эллиптированной) информации и/или степени статичности vs. динамичности текста). В ситуации речевой коммуникации тип и способ взаимодействия коммуникативной и смысловой структур в существенной степени зависит от цели и задачи коммуникантов (ср. Падучева 2001). В общем случае **неединственность** коммуникативного структурирования определяется структурой коммуникативной ситуации. На такого рода неединственность обратила в свое время внимание Е. В. Падучева, отметив, что различие в описании коммуникативной структуры высказывания — это различие в его функционировании при решении тех или иных задач (Падучева 2001: 109–112).

Поэтому в акте успешной коммуникации говорящий/пишущий и слушатель/читатель могут определять границы тем и рем по-разному; эти интерпретации являются одинаково правильными.

При исследовании звучащего текста важно, что говорящим актуализируется некоторый набор возможных вариантов членения. Этот набор является (в идеальном варианте правильного чтения) подмножеством вариантов коммуникативного членения письменного текста. В случае звучащего текста в качестве показателей темы и/или ремы выступают дополнительные (интонационные) признаки, они указывают на более высокий или низкий статус актуализации темы.

Кредо автора настоящей работы отличается от кредо некоторых исследователей актуального членения и, в частности, Т. М. Николаевой (Николаева 2000: 70)[3]: я солидаризируюсь с позицией В. Б. Касевича, согласно которой «тема как предмет сообщения противополагается «всему остальному», а «все остальное», в свою очередь может быть суждением» (Касевич 2006). По-видимому, именно такое кредо вытекает и из когнитивного, и из прикладного подходов к исследованию восприятия текста.

---

[2] Именно поэтому традиционно принятый термин «актуальное членение предложения» оказывается более чем спорным в случае анализа структурных составляющих связного текста, т. к. теряется представление о связности текста. Термин «предложение» в таком исследовании оказывается неудачным, т. к. лишает акцента на текстовый контекст, который и актуализирует тот или иной вариант коммуникативного членения.

[3] Ср. соотнесение акцентного выделения с компонентом ремы (Николаева 1982).

Наиболее существенными для данной работы являются следующие положения:

a) Важным является выделение Т. Е. Янко двух основных типов коммуникативных значений: конституирующее речевой акт и не-конституирующее (подробнее см. Янко 2001). Естественно, **конституирующий компонент является обязательным**, а не-конституирующий компонент — факультативным. Традиционно в теории актуального членения рассматривают речевой акт как реализацию сообщения; в этом случае рема — это конституирующий компонент, а тема — не-конституирующий. В данной работе для любых типов речевых актов конституирующий компонент будем называть ремой, а не-конституирующий — темой.
b) «В языках существует не менее четырех способов маркирования темы, на которые может опираться воспринимающий речь человек. Это позиционный, грамматический, лексический и фонетический способы. Первый и последний из них, можно думать, являются универсальными, в то время как остальные два представлены в одних языках, но отсутствуют в других» (Касевич 2006: 600).
c) Наряду с бинарным противопоставлением тема vs. рема **может** рассматриваться и шкалирование.

В данной статье на материале делового и художественного текстов, уже обсуждаемых ранее (напр., (Ягунова 2007а, Ягунова 2007б)), я хочу обсудить основные принципы экспериментального определения коммуникативного членения. В деловом тексте доминирует информативная функция языка, его высказывания рассматриваются как реализация сообщения, то есть материал в наибольшей степени соответствует стандартной платформе для исследования коммуникативного членения высказывания, в нем преобладает нормативная реализация и наличие обоих компонентов («тема»+«рема»). Для художественного текста картина существенно другая. Сюжетный текст с элементами диалога не ограничивает множество своих высказываний: кроме сообщений, в нем представлены и вопросы, и императивы, и восклицания (не говоря уж о компонентах, вводящих прямую речь). За соответствующими коммуникативными структурами ***всех*** этих высказываний закрепляем терминологию «тема» и «рема». Многообразие этих структур заставляет предположить, что не все из них являются расчлененными (или эксплицитно расчлененными), то есть факультативный компонент темы в них может быть эксплицитно не выражен.

В статье рассматриваются возможности коммуникативного членения высказываний текста:

- в звучащей vs. письменной форме;
- для текстов делового vs. художественного функционального стиля.

В исследовании используется три разных экспериментальных подхода:

- экспертный (группа экспертов-лингвистов) на материале звучащих текстов,

- экспериментальный (группой наивных носителей языка) на материале письменных текстов,
- вычислительный: с применением алгоритма И. В. Азаровой и С. Гордеева (на материале письменных текстов).

Может ли исследователь **сам** осуществить коммуникативное членение? Да, но в ограниченных пределах. Для решения многих задач анализа текста, понимания и/или извлечения знаний знание конечной задачи исследования может препятствовать адекватности самостоятельного решения задачи коммуникативного членения. Именно поэтому для нас так актуальны варианты экспериментальных решений поставленной задачи.

Идеи, лежащие в основе данной работы:

1. Функциональный стиль и тип текста определяет коммуникативное членение высказываний текста.
2. Рассматриваемые три экспериментальных подхода для определения коммуникативного членения соотносятся с разными задачами:
   - первый подход позволяет получить вариант членения высказываний текста (а) максимально независимый, прежде всего, от установок экспериментатора и цели работы; (б) учитывающий структуру текста в целом;
   - второй подход позволяет получить вариант членения высказываний текста (а) максимально учитывающий модельные установки эксперимента, прописанные в инструкции; (б) не учитывающий структуру текста в целом;
   - третий подход позволяет получить вариант членения высказываний текста (а) максимально независимый, прежде всего, от установок экспериментатора и цели работы; (б) максимально учитывающий формально-синтаксические и позиционные параметры; (в) не учитывающий структуру текста в целом.

Каждый из трех экспериментальных подходов имеет свои преимущества и недостатки; вероятно, отдельные данные некоторых экспериментов будут выглядеть несколько странно (см. табл. 3 и 4), однако это не мешает интегральной оценке коммуникативной структуры текстов.

2. Материал. Методика

В качества материала исследования использованы:

- отрывок из официальной публикации «Закон об иностранных инвестициях во Вьетнаме и нормативные акты, изданные на его основе» (в дальнейшем, деловой текст);
- отрывок художественной вьетнамской прозы Нам Као «Ти Фео» с элементами диалога (в дальнейшем, художественный текст).

Тексты рассматриваются как в письменной форме представления (эксперимент 2 и 3), так и в звучащей (в чтении опытного диктора) — эксперимент 1.

2.1. Дизайн экспертной оценки для определения коммуникативного членения высказываний звучащего текста профессиональными лингвистами — ЭКСПЕРИМЕНТ 1

В качестве первого подхода рассматривается экспертная оценка группой лингвистов-экспертов, которые могут (1) осуществлять коммуникативное членение высказывания в контексте текста; (2) основываться на существенно большем репертуаре возможных оснований для решения задачи выделения темы; (3) справляться с конкуренцией критериев. Адекватность результатов нашего эксперимента обусловливается представительностью, уровнем квалификации и сбалансированностью группы экспертов-лингвистов (13 человек), в которую входят фонетисты и специалисты в области семантики и синтаксиса.

При выборе такого экспериментального дизайна в случае первого варианта можно столкнуться с артефактами, порожденными особенностями научных взглядов лингвистов-экспертов. Однако подбор сбалансированной — с точки зрения научных предпочтений — группы делает такого рода помехи минимальными.

Экспертам-лингвистам было предложено неограниченное число раз прослушать исходные тексты в удобном режиме и в орфографической записи подчеркнуть слова, относящиеся к теме (в терминах актуального членения предложения). В инструкции подчеркивалась необходимость ориентироваться в решении поставленной задачи на чтение диктора, диктор в своем прочтении актуализировал варианты коммуникативного членения.

2.2. Дизайн эксперимента по определению коммуникативного членения высказываний письменного текста наивными носителями языка — ЭКСПЕРИМЕНТ 2

В качестве второго подхода рассматривается эксперимент с письменным текстом, в котором участвует 56 наивных носителей языка, студентов ПНИГУ[4]. Таким образом, мы избегаем тех артефактов, которые были возможны в эксперименте 1. Однако причины других артефактов заложены

[4] Пользуясь случаем, хочу поблагодарить Е. В. Ерофееву, Т. И. Доценко и пермских студентов за помощь в проведение эксперимента 2.

в самой инструкции[5], т. к. по условиям такого эксперимента испытуемые не должны иметь исходного представления об актуальном членении предложения (см. Грудева 2007; Грудева 2001). В письменном тексте заложена бОльшая неоднозначность коммуникативного членения.

Итак, вот уже опробованная в докторском исследовании Е. В. Грудевой *инструкция по определению темы высказывания*: «В лингвистике принято выделять тему и рему в предложении. Тема — это слово (или сочетание слов), обозначающее явление, предмет и т. п., которому посвящено предложение (о котором говорится в предложении). Напр., в предложении *Петя — студент* темой является слово «Петя» (т. е. в этом предложении говорится о Пете). В предложении *Волноваться вредно* темой является слово «волноваться» (т. е. в этом предложении говорится о состоянии волнения). В предложении *Гильдия киноактеров США объявила номинантов на свои профессиональные премии* темой является сочетание слов «гильдия киноактеров США» (т. е. в этом предложении говорится о гильдии киноактеров США). В предложении *В поле работает трактор* темой является словосочетание «в поле» (т. е. в этом предложении говорится о поле). Ваша задача заключается в том, чтобы подчеркнуть в предложениях текста тему» (Грудева 2007: 228).

2.3. Дизайн вычислительного метода определения коммуникативного членения высказываний письменного текста — ЭКСПЕРИМЕНТ 3

Идея эксперимента И. В. Азаровой и С. С. Гордеева описана в статье (Азарова, Гордеев 2007). Для меня это некоторая формальная процедура, осуществленная И. В. Азаровой и С. С. Гордеевым для рассматриваемых мной текстов, во-первых, дополняющая систему правил описания коммуникативной структуры введением дополнительных компонентов; во-вторых, имеющая право на отказ (отсутствие разбора)[6].

В дополнение к дихотомическому описанию тема-рематических структур И. В. Азарова и С. С. Гордеева вводят «внерематические» компоненты, следуя идеям И. Ф. Вардуля (1989) (ср. Гордеев, Азарова 2007):

- Ситуативы (S) — определительные или обстоятельственные обороты, которые относятся к ситуации в предложении в целом и задают своеобразное условие, в котором реализуется противопоставление темы-ремы;
- конструкция — обычно предикат или субъект+предикат, не являющиеся темой или ремой предложения, а представляющие нек. лицо и его

[5] В предлагаемой информантам инструкции может быть и причина нарочитой «модельности» эксперимента, если инструкция провоцирует строго определенные (заданные экспериментатором) реакции информантов.

[6] Пользуясь случаем, хочу поблагодарить И. В. Азарову и С. С. Гордеева за проведение эксперимента 3 и предоставление мне всех материалов.

отношение (в широком смысле) к предмету сообщения — (Р): *мы выяснили/ подсчитали…, Полинг и Кюри установили…*,
- Пояснения к теме или (чаще) реме (Е), уточняющий компонент или перечень значений (обычно дается в скобках);
- Причинно-следственные и другие логические связки (L), осуществляющие внутритекстовую связанность.

3. Результаты. Описание тема-рематической структуры высказывания и функциональный стиль текста

Даже результаты эксперимента 2 (в данном случае наиболее непредсказуемого, т. к. элементы темы выделялись наивными носителями языка) демонстрируют сравнительно высокую согласованность: все данные укладываются в интервал ±2д.

В отличие от художественного текста (особенно от реплик диалога) деловой текст имеет сравнительно четкую тема-рематическую структуру: в высказываниях делового текста присутствуют слова, отвечающие теме, и испытуемые или автомат не испытывают затруднений в их выделении. По результатам первых двух экспериментов все слова текста были отнесены к одному из следующих компонентов (названия групп условные): ядро темы (Тя), периферия темы (Тп) и нетема (неТ) (см. табл. 1). Ядро темы представляли слова, выделенные более чем 69% информантов, периферию темы — слова, выделенные более чем 38% информантов и менее, чем 69% (промежуточная группа), нетема — слова, выделенные менее чем 38 информантами. Таким образом, «группа темы» в предложениях текста рассматривается как неоднородная, включающая «ядро» и «периферию».

Таблица 1. Объем компонентов в абсолютном и относительном числе словоупотреблений для экспериментов 1 и 2

		Тя (> 69% ии.)		**Тп (> 38% и < 69% ии.)**		**неТ (< 38% ии.)**	
Эксперимент	**стиль текста**	**абс**	**%**	**абс**	**%**	**абс**	**%**
Эксперимент 1 (звучащий)	деловой	19	8	**92**	**40**	**117**	**51**
	художественный	13	6	42	19	**166**	**75**
Эксперимент 2 (письменный)	деловой	9	4	11	5	**208**	**91**
	художественный	2	1	39	18	**108**	**81**

По результатам третьего эксперимента все слова текста были отнесены к одному из компонентов «усложненной структуры»: теме, реме, а также «внерематических и/или внетематических компонентов» (см. табл. 2, условные обозначения см. выше).

Таблица 2. Объем компонентов в абсолютном и относительном числе словоупотреблений для эксперимента 3[7]

	T		R (неТ)		прочие								отказ	
					S		P		E		L			
стиль текста	**абс**	**%**	**абс**	**%**	**абс**	**%**	**абс**	**%**	**абс**	**%**	**абс**	**%**	**абс**	**%**
деловой	77	37	76	37	43	21	3	1	29	14	4	2	0	0
художественный	18	8	60	27	19	9					4	2	119	39

4. Обсуждение результатов. Выводы

Для делового текста разделение словоупотреблений лингвистами-экспертами на элементы темы и элементы ремы (нетемы) соответствуют очень четким синтаксическим структурам высказываний текста (см. рис.1А). Выделение ядра и периферии в целом соотносится с построением синтаксической структуры, где ядру соответствовали вершины синтаксических конструкций, а периферия, как правило, представляла собой определяющие и уточняющие части этих конструкций. Результаты коммуникативного членения по данным эксперимента 1 и 3 (а во многом и по данным эксперимента 2) демонстрируют существенную согласованность (см. табл. 3), хорошо интерпретируемую благодаря введению дополнительных компонентов.

Каждое предложение делового текста содержит в себе хотя бы одно слово, относящееся к ядру темы, и ряд слов, относящихся к периферии темы (см. табл. 3). Для художественного текста подобное утверждение неверно.

Для художественного текста выделение темы даже лингвистами-экспертами представляло серьезную проблему. Элементы темы в тексте крайне редки (см. табл. 4): как правило, предшествующий контекст определяет «о чем говорится в предложении», то есть можно условно сказать, что тема большинства высказываний не выражена непосредственно в этом высказывании (имплицитна). Среди элементов темы основное место занимают обозначения действующих лиц.

В художественном тексте высокая доля нерасчлененных коммуникативных структур (см. табл. 4). Следующие признаки коммуникативного структурирования художественного текста хорошо коррелируют с универсалией Гивона (Givyn 1983):

- наличие большого количества клауз (и даже высказываний) без компонента «тема»;
- отнесение к компоненту «тема» слов, легко восстанавливаемых на основании предшествующего контекста, прежде всего, местоимений.

[7] S, P, E(R), L — условные обозначения для дополнительных компонентов п. 2.3. (Гордеев, Азарова 2007).

Согласно этой универсалии, чем больше активация референта, тем больше вероятность того, что он будет выражен более редуцированно.

В автомате — эксперимент 3 — вряд ли могут быть хорошо прописаны правила принятия решения в условиях большого количества имплицитной информации, рубленных (включая элементы эллипсиса) высказываний. Даже в случае эксперимента 2 правила, предписываемые инструкцией, не могут охватить многообразия структурных вариантов, а некоторые результаты плохо интерпретируемы: напр., «*Забирай* (неТ) */ и катись* (Тп)», «*– Ладно, ладно*, (неТ) */ бери* (Тп)…», *Впервые слышу* (Тп) */ от тебя такое* (неТ)!» и т. д. Но эти спорные случаи не мешают интегральной оценке.

Данные, представленные в табл. 3 и 4, позволяют интегрально взглянуть на представление о коммуникативных структурах анализируемых текстов: данные каждого из экспериментов — это результаты особых экспериментальных подходов. Даже в том случае, если какой-либо из вариантов членения по одному из экспериментов вызывает сомнения, он все равно выполняет функцию выделения сбоя работы процедур. В табл. 3 и 4 фрагменты текста сгруппированы в соответствии с данными по всем трем экспериментам. Сочетания слов (иногда достаточно пространные), для которых результаты совпадают, занимают единую ячейку. Слова и сочетания, попавшие в одну ячейку, мы рассматриваем как особые коммуникативные структурные составляющие (КСС) текста. Для делового текста — и во многих случаях для художественного текста — эти КСС не противоречат правилам синтаксического структурирования.

Тематические КСС, выделенные на основании всех трех экспериментов (в первых двух экспериментах как ядро темы), соотносятся с основными ключевыми словами этих текстов, они же маркируют композиционные фрагменты текста и их темы (см. табл. 4, выделение кеглем). В деловом тексте в качестве максимально актуализированных ядер темы выступают ключевые и опорные слова соответствующей предметной области (*стороны, плановые показатели, заявка, импортные квоты* и т. д.), подробнее см. (Ягунова 2007б.). В художественном тексте в этой роли явно выступает лишь одно из действующих лиц (*Советник*).

В ряде случаев для художественного текста КСС объединяют элементы из разных предложений (но одной ситуации); это синтаксически спорный случай объединения, он относится скорее к структурированию ситуации, особенно на пиках эмоциональной напряженности (в табл. 4 такие ячейки имеют диагональное перечеркивание).

Интонационные характеристики могут — для эксперимента 1 — актуализировать определенный вариант структурирования (Ягунова 2008)[8]:

- для ***делового*** текста ***пауза*** в большинстве случаев маркирует границу между компонентами темы и ремы;
- для ***художественного*** текста складывается противоположная картина: ***пауза*** крайне редко маркирует рассматриваемую границу, а ***мелодическое выделение*** (движение основного тона более чем на 25 Гц) маркирует

[8] Подробнее см., например, (Halliday 1970; Brown 1983; Terken 1980; Касевич 2006).

новое (Ягунова 2007а), происходит взаимодействие коммуникативной и смысловой структур.

По-видимому, паузы (как границы между группами темы и ремы) не являются существенным фактором (см. табл. 3). А вот наличие мелодического выделения, включающего границу высказывания, вероятно, может организовывать кульминационную ситуацию для художественного текста (см. в табл. 4 ячейки с диагональным перечеркиванием). Впрочем, это объединение выявляется даже по материалам эксперимента 2 (несмотря на инструкцию). Для эксперимента 3 эти фрагменты оказываются недоступными для разбора по правилам («отказ»).

Таблицы 3 и 4 содержат анализируемый текст (включая варианты членения), именно поэтому они приводятся целиком. Таблицы требуют усилий для чтения, но в них представлена максимальная интерпретация (см. условные обозначения). В таблицах 3 и 4 приводятся также данные о нетематических составляющих, однако их интерпретация — предмет отдельной статьи.

Таблица 3. Пример отнесения словоупотреблений делового текста к компонентам коммуникативного членения на основании данных экспериментов 1–3[9]

Словоупотребления	**Э.1**	**Э.2**	**Э.3**
Стороны,	Тя	Тя	Т
~~*участвующие в реализации*~~	неТ	неТ	Т
договора о деловом сотрудничестве	Тп	неТ	Т
и предприятия	Тя	неТ	Т
с иностранным капиталом	Тп	неТ	Т
~~**имеют право самостоятельно устанавливать программы и планы своей производственно- хозяйственной деятельности**~~	неТ	неТ	неТ
Плановые показатели	Тя	Тя	Т
вьетнамскими государственными органами	неТ	неТ	Р
~~**не устанавливаются**~~	неТ	неТ	неТ

[9] Условные обозначения к таблицам 3 и 4:
п/ж шрифт — отнесение на основании результатов трех экспериментов;
курсив — отнесение на основании результатов двух экспериментов;
увеличенный кегль — выделение ядерных тематических составляющих на основании результатов 3 экспериментов;
отсутствие перечеркивания — тематическая составляющая (ядро или периферия);
наличие перечеркивания — нетематическая составляющая;
ячейки с диагональным перечеркиванием — составляющая включает границу высказывания.

Таблица 4. Пример отнесения словоупотреблений художественного текста к компонентам коммуникативного членения на основании данных экспериментов 1–3

Словоупотребления	**Э.1**	**Э.2**	**Э.3**
Когда одолевают подобные	Тп	неТ	S
мысли	Тя	Тп	S
~~*трудно сохранить хорошее*~~	неТ	неТ	S
настроение	неТ	Тп	S
~~*и выдержку*~~	неТ	неТ	S
~~*А тут еще*~~	неТ	неТ	L
~~**принесла нелегкая Ти Фео**~~	неТ	неТ	неТ
~~**не иначе как явился**~~	неТ	неТ	неТ
~~*клянчить*~~	неТ	Тп	неТ
~~**на выпивку**~~	неТ	неТ	неТ
Советник	Тя	Тя	Т
~~**решил бросить ему пять хао, чтобы поскорее убрался, но при этом не выдержал и крикнул:**~~	неТ	неТ	неТ
~~*Опять*~~	неТ	неТ	
притащился	неТ	Тп	
~~*Пора бы знать свое*~~	неТ	неТ	
место	неТ	Тп	
~~*Здесь тебе не банк*~~	неТ	неТ	

Заключение

Интегральная оценка коммуникативного членения, опирающаяся на результаты всех трех экспериментов, позволяет объемно взглянуть на природу коммуникативного членения структурных составляющих текста и текста в целом. Даже те оценки, которые плохо интерпретируемы, важны как показатели потенциальных сбоев. Тематические КСС, выделенные на основании всех трех экспериментов, соотносятся с основными ключевыми словами этих текстов, они же маркируют композиционные фрагменты текста и их темы. Наибольший интерес представляют экспериментально выделяемые КСС как мало изученный тип составляющих. Для делового текста — и во многих случаях для художественного текста — КСС не противоречат правилам синтаксического структурирования. Особые КСС — в случае художественного нарратива — могут относиться скорее к структурированию ситуации, а не синтаксической структуры, что может подчеркиваться как интонационные кульминационные моменты (ситуации).

Литература

1. *Вардуль И. Ф.* 1989 — О языковых типах в параметре порядка слов // Очерки типологии порядка слов. М., 1989.
2. *Гордеев С. С., Азарова И. В.* 2007 — Характер корреляции между порядком слов и коммуникативной перспективой в научно-технических текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии. М., Наука, 2007. — С.130–136
3. *Грудева Е. В.* 2001 — Тематические элементы и мера избыточности текста (на материале русского языка в письменном варианте) // Язык и речевая деятельность. 2001. Т. 4, ч. 1. –СПбГУ, 2001. — С. 161–172.
4. *Грудева Е. В.* Избыточность и эллипсис в русском письменном тексте — Череповец, 2007.
5. *Касевич В. Б.* 2006 — Семантика. Синтаксис. Морфология // Труды по языкознанию. СПб, 2006
6. *Касевич В. Б.* 2009. — Топик: понятийная категория? Научное наследие Владимира Григорьевича Адмони и современная лингвистика — Материалы Международной научной конференции, посвященной 100-летию со дня рождения В. Г. Адмони (9–13 ноября 2009 года). СПб.: «Нестор-История», 2009, с. 127–128
7. *Николаева Т. М.* 1982 — Семантика акцентного выделения. М., 1982.
8. *Николаева Т. М.* 2000 — От звука к тексту. Москва, Языки русской культуры. 2000
9. *Падучева Е. В.* 2001 — Высказывание и его соотнесенность с действительностью М., 2001
10. *Хоккет Ч.* 1970– Проблема языковых универсалий // Новое в лингвистике. Вып. V. — М.: Иностр. лит., 1970. — С. 45–76
11. *Ягунова Е. В.* 2007а — Тема и рема, данное и новое при восприятии текста // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» (Бекасово, 30 мая — 3 июня 2007 г.) // Под ред Л. Л. Иомдина, Н. И. Лауфер, А. С, Нариньяни, В. П. Селегея. — М.: Издательский центр РГГУ, 2007, с. 610–617.
12. *Ягунова Е. В.* 2007б — Коммуникативная и смысловая структуры текста и его восприятие // Вопросы языкознания, 2007, 6. — С. 32–49
13. *Ягунова Е. В.* 2008 — Вариативность стратегий восприятия звучащего текста (экспериментальное исследование на материале русскоязычных текстов разных функциональных стилей). Пермь, 2008
14. *Янко Т. Е.* 2001 — Коммуникативные стратегии русской речи. М., 2001
15. *Brown G.* 1983 — Prosodic Structure and the Given/New Distinction. // Prosody: Models and Measurements. Ed. by A. Cutler and D. R. Ladd. Berlin e. a., 1983.
16. *Givón, Talmy* 1983. — Topic continuity in discourse: An Introduction // Topic continuity in discourse: a quantitative cross-linguistic study (Typological studies in language, vol. 3). — Amsterdam: Benjamins. 1983
17. *Halliday M. A. K.* 1970 — A Course in spoken English: Intonation. Oxford, 1970.
18. *Terken J. M. B.* 1980 — The Distribution of Pitch Accent in Descriptive Language as a Function of Informational Variables. // IPO Annual Progress Report. 1980, № 15.

EXTRACTING SUBJECTIVE INFORMATION FROM SPOKEN DISCOURSE: THE CASE FOR PROSODIC EMPHASIS[1]

Yanko T. E. (tanya_yanko@list.ru)

The Institute for Linguistics, Russian Academy of Science, Moscow, Russia

The aim of this paper is to develop a method for extracting subjective information from spoken discourse. The example of subjective information discussed here is the attitude of the speaker to the reported events as to rare or abnormal. Disrupting a routine course of life which causes the speaker to express his/her strong feelings refers to the concept of emphasis. Emphasis basically has a prosodic expression. The emphatic prosody designates: 'it is great that he did it by himself because generally he needs assistance', 'he came to the meeting which is very unusual because he never comes', 'it was the president who congratulated them which is highly honorable because generally it is the vice-president who does it'. Instances of emphasis are highly frequent in everyday speech as well as in mass media language, especially in breaking news. Reports of achievements and losses, accidents, severe meteorological phenomena, such as hurricanes, earthquakes, swollen streams, and numerical scores of any kind are generally made with the use of emphatic prosody. The instrumental analysis of the physical parameters of emphasis can serve as a means for extracting subjective information from spoken texts. The results presented here are exemplified by records from oral corpora. The frequency tracings are generated by the software program Speech Analyzer.

Key words: subjective information, spoken discourse, prosody, emphasis, communicative structure

The aim of this paper is to develop a method for extracting subjective information from spoken discourse. The example of subjective information discussed here is the attitude of the speaker to the reported events as to rare or abnormal. Disrupting a routine course of events which causes the speaker to be surprised and to express his/her strong feelings about it refers to the concept of emphasis. Emphasis basically has a prosodic expression. The emphatic prosody designates: 'it is great that he did it by himself because generally he needs assistance', 'he came to the meeting which is very unusual because

[1] This work has been supported by the Program of fundamental research "Language and literature in the context of social dynamics", 2012–2014, section "Electronic databases and multimedia technologies in Philology", the Russian Academy Presidium, Department of language and literature, project "The computer model and electronic data base "Meaning-Spoken Speech"" [Rabota vypolnena po Programme fundamental'nyh issledovanii Presidiuma RAH, sektsii iazyka i literatury «Iazyk i literatura v kontekste kul'turnoi dinamiki, 2012–2014, razdel «Elektronnyie bazy dannyh i mul'timediinyie tehnologii v filologicheskih issledovaniiah », proekt «Razrabotka kompiuternoi modeli "Smysl-Zvuchashchaia rech" i elektronnoi bazy dannyh dl'a ee podderzhki »].

he never comes', 'it was the president who congratulated them which is highly honorable because generally it is the vice-president who does it'. Instances of emphasis are highly frequent in everyday speech as well as in mass media language, especially in breaking news. Reports of achievements, victories, losses, accidents, severe meteorological phenomena, such as hurricanes, earthquakes, swollen streams, and numerical scores of any kind are generally made with the use of emphatic prosody. The instrumental analysis of the physical parameters of emphasis can serve, therefore, as a means for extracting subjective information from spoken texts. The results presented here are exemplified by records from the oral corpus specifically targeted at analyzing emphasis. The corpus (≈ 1 hour 15 minutes of sounding) contains records of everyday speech and breaking news from TV and radio broadcasting, as well as experimental laboratory readings with the forecasted parameters of emphasis. The waveforms and frequency tracings of examples considered below are generated by the software program Speech Analyzer.

Henceforth, the basic aim of this paper is to consider the semantic properties and the prosodic means of emphasis. The point of departure for analyzing emphasis is Russian. The contrastive analysis of Russian, English, German, French and Danish data demonstrates a significant similarity in prosodic means of emphasis: the examined languages display very similar cues for emphasis.

The semantics of emphasis is to express the speaker's attitude with respect to abnormal phenomena of life. In the question *Do you actually have a son?!*, the speaker, first, is making sure whether the hearer really has a son, and, secondly, expresses the idea that the hearer is too young to have a son and that if it is true it must make the hearer be proud of it. Emphasis, therefore, refers a current state of affairs to some ordered set where all elements are ranged in accordance with a certain standard of life. The emphasized element of the semantic structure is always the most unusual or one of the most unusual performers of the part it plays. It is placed close to the extreme of probability scale with the minimum value.

The concept of emphasis is a communicative meaning, and its basic property is the ability to combine with other communicative meanings. As a result, emphatic topics, emphatic foci, and emphatic questions arise. Therefore, emphasis does not function in a sentence independently. It can only combine with the meanings which form the components of speech acts: either statements or questions.

The principal — but not the only one — prosodic parameter of emphasis can be described as follows. The basic tone — the falling one which originally designates a focus or the rising one which designates a topic or a *yes-no*-question — in the context of emphasis associates with a tone which precedes the basic one and points in the opposite direction. Hence, with emphasis, the original fall is being preceded by a rise, while the rise is being preceded by a fall. This prosodic phenomenon might be called "emphatic tonal curving". To consider this point we need an example.

The prosodic parameters of emphasis are instantiated here by laboratory readings of the Russian word *sam* 'himself' (masculine) whose lexical semantics agrees with the concept of communicative emphasis (cf. [Yanko 2001: 287–289]), because *sam* means that what the agent unexpectedly did by himself s/he used to do with the help of his/her assistants or substitutes. In example (1) the word *sam* 'himself' is employed in a variety of communicative functions, either composed with emphasis or not: a) *Sam*? 'Did

he do it by himself?' (a non-emphatic *yes-no*-question), b) *Sam*?! 'Did he really do it by himself?! Unbelievable!' (an emphatic *yes-no*-question), c) *Sam* 'He did it by himself' (a non-emphatic statement), d) *Sam*! 'He did it by himself. It is unbelievable!' (an emphatic focus).

(1) a) *Sam*?; 'himself?' b) *Sam*?!; 'himself?!' c) *Sam*; 'himself', d) *Sam*! 'himself!'

In minimal pairs (a)-(b) and (c)-(d) respectively the second member of each pair is an emphatic counterpart of the first one. The pairs exemplify composing communicative meanings (or functions) — questions, topics, and foci — with emphasis. These examples demonstrate how emphasis modifies the fundamental frequency course when it combines with the communicative meanings originally designated by rises or falls of frequency. This description is obviously not complete without the pair "topic vs. emphatic topic" which is omitted on the assumption that topics and *yes-no*-questions, either composed with emphasis or not, are expressed identically. In fact, there is some difference between the expression of topics and *yes-no*-questions but here it is ignored: the analysis presumes that both topics and *yes-no*-questions in Russian are designated by the rise of the fundamental frequency which is placed on the stressed syllable of the accent-bearer of a topic or a question. The "emphatic" counterpart of a rise is therefore a rise which is preceded by an additional fall of frequency or by a low level tone, cf. frequency tracings of examples (1a) and (1b) below. Similarly, the difference between (1c) and (1d) is an additional "emphatic" rise of frequency in (1d) which precedes a fall as contrasted to (1c) where a fall designates a simple focus.

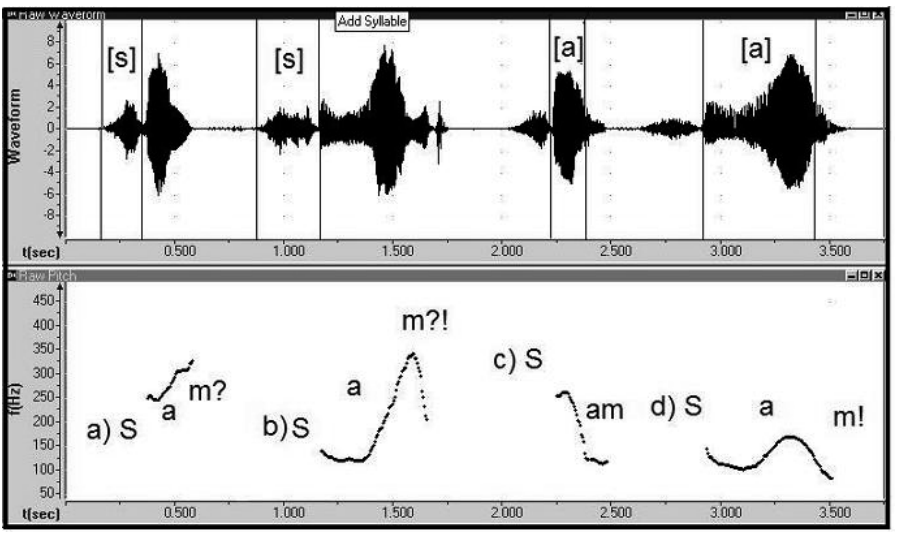


Fig. 1. The waveform and frequency tracing of Example (1)

The frequency tracing presented in Figure 1 on the lower panel, indeed, demonstrates "tonal curving" in (b) as contrasted to (a) and in (d) as contrasted to (c).

Nevertheless, the tonal "curving" is not the only cue for emphasis. The frequency tracing shows that emphatic question (1b) has wider pitch-range than non-emphatic question (1a), namely — the rise in (1b) has a lower starting point than the initial frequency in (1a) and a higher peak frequency than the one in (1a). The accented — and the only one here — syllable of the emphatic accent-bearer has a significantly longer duration in the emphatic instances of *sam* than in the non-emphatic ones. Besides, in the context of emphasis the rise (either in a topic or in a question) or a fall (in a focus) is delayed because of a longer duration of the level onset of the syllable, especially because of the duration of the initial consonant. In example (1) it is [s]. The waveform presented by the upper panel (cf. Figure 1) shows the contrastive duration of the initial [s] in non-emphatic question (1a) and in emphatic question (1b). The intercepts of the waveform referring to [s] in (1a) and (1b) are delimited by vertical tracking.

The waveform also demonstrates the suppression of amplitude at the onset of the syllable, cf. vowel [a] in non-emphatic (1c) and emphatic (1d). Amplitude is rather low at the onset of a syllable but from approximately the middle of a syllable it grows sharply. The intercepts of the waveform referring to [a] in (1c) and (1d) are marked by vertical tracking.

An emphasized syllable gets, therefore, a binary structure: the delayed frequency peak and suppressed amplitude divide a syllable into two parts. As for the overall duration of the accented syllable of the accent-bearer, with emphasis it is about twice the corresponding non-emphatic one.

The binary structure of emphatic prosody is exemplified below by English and French emphatic sentences (5)–(6) with polysyllabic accent-bearers which in the context of emphasis get a secondary stress at the onset of a word. The components of the bidirectional emphatic frequency contours are, therefore, placed on the initial and the terminal accented syllables of the accent-bearer each. About emphasis in French cf. [Féry 2001] and [Faustova 2006].

While laboratory example (1) demonstrates the complete set of "emphatic" parameters, the examples taken from spontaneous speech do not always keep all correlations in full. It might be hypothesized that feelings of the speaker expressed by prosodic emphasis are being measured by the speaker depending on his/her agitation level. Nevertheless, all examples presented below where emphasis keeps its identity maintain the tonal curving. Tonal bending to the direction opposite to the original direction of the topical, focal, or interrogative prosodic cues may be positioned on the initial part of the accented vowel of the accent-bearer or even on a prolonged initial voiced consonant as in Russian sentence (2).

Sentence (2) exemplifies a speech of a lawyer whose client was adjudged dead by his ex-wife and was not able to restitute his rights for a year. In the emphatic topic *bolee goda* 'more than a year' the fall of frequency which precedes the "topical" rise is placed on the consonant [g] in the word *goda* 'of a year' which has a significantly longer duration than the one in the second — non-emphatic — occurrence of the same word: in order to be more convincing the lawyer pronounces the fragment *bolee goda* twice.

(2)	*I*	*vot*	*uzhe*	*bolee*	*goda,*	*bolee*	*goda…,*
	And	here	already	more than	a year	more than	year…

<on ne mozhet vosstanovit' svoe polozhenie kak fizicheskoe litso.>
<he is not able to restitute his position as a physical person>.

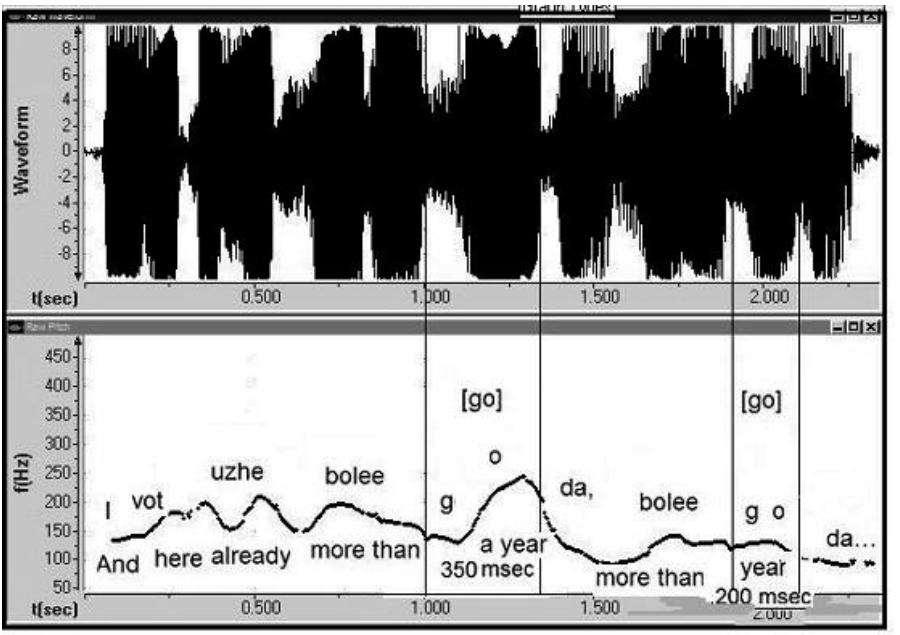


Fig. 2. The waveform and frequency tracing of Example (2)

The accented syllable [go] of the accent-bearer *goda* 'of a year' in the emphatic topic *I vot uzhe bolee goda* 'And now already more than a year' has a duration of 350 msec while the syllable [go] within the second non-emphatic occurrence of the same word only lasts 200 msec. Two occurrences of the same word successfully allow for contrasting the emphatic and the non-emphatic sounding. Emphasis in this sentence has longer duration, larger range of frequency, and syllable-initial lowering the amplitude of oscillation.

Example (3) illustrates, for comparison, the Danish emphatic question with the meaning of 'unbelievable'. The speaker expresses his surprise with respect to the fact that the voice belongs to Jens and not to Lena: — *Who?! –Oh, Jens. <Well, I thought it was Lena>*. Consider the full text of a dialogue.

(3) (a) — *Hej*! *Det er Jens i telefonen*.
'High! It's Jens'.
(b) — *Hvem er det*?!
'Who is this?!'
(c) — *Jens*.
'Jens'.
(d) — *Ah, det er Jens*?! [a deɐ jens]
'Oh, this is Jens?!'
(e) — *Jeg troede ellers, det var Lene*.
'I believed it was Lena'.

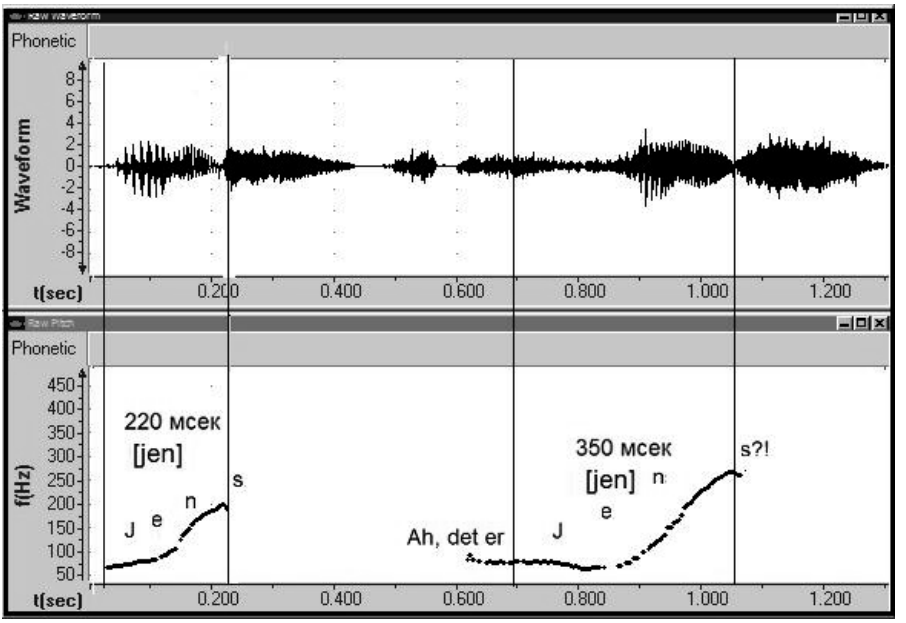


Fig. 3. The waveform and frequency tracing of Example (3c-d)

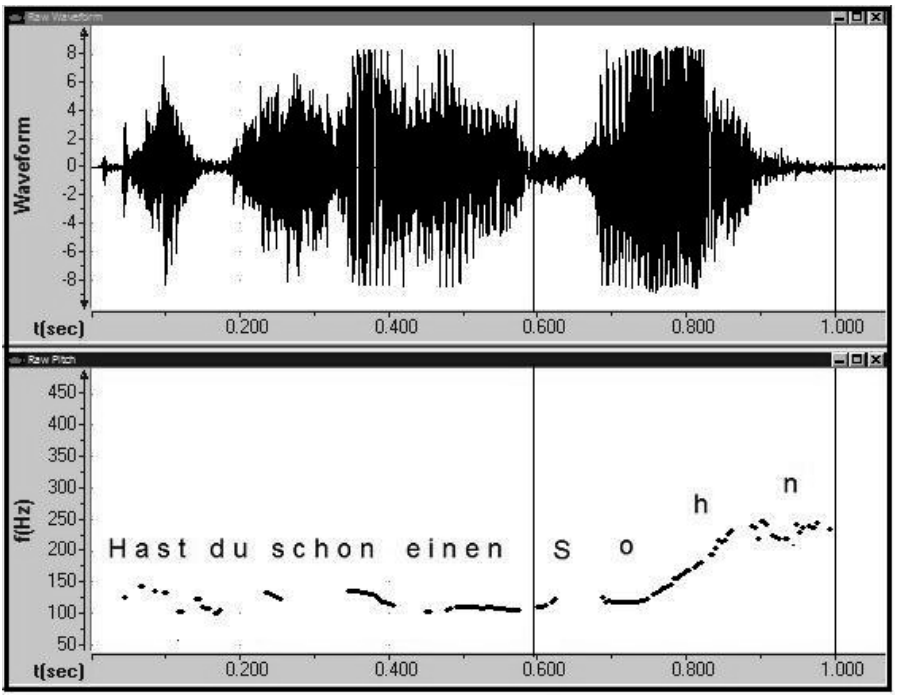


Fig. 4. The waveform and frequency tracing of Example (4)

In example (3) the emphatic and non-emphatic rises are displayed by the same segmental material, namely — the word *Jens*. The waveform and frequency tracing of example (3c-d) demonstrates the amplitude increase, especially its higher peak value, a longer duration of the emphatic syllable, and the amplitude suppression within the consonant [j] in the emphatic sentence (3d) as compared with the non-emphatic preceding reply (3c).

German emphatic question (4) also displays all properties of emphasis. The frequency tracing demonstrates "low level — rising" course of frequency within the accent-bearer the word *Sohn* 'son'. The waveform displays the binary structure of the vowel [o], and a longer duration of the word-initial consonant [z].

(4) *Hast du schon einen Sohn*?!
'Do you already have a son?!'

The binary structure of the cues for emphasis allows for positioning the emphatic contour on the two syllables of a word with a secondary stress: the first component of a "curved" frequency course is placed on the first stressed syllable of a word with a secondary stress, the second — the basic — component is placed on the main word stress. Emphasis can, hence, capture two syllables of a polysyllabic word.

In English example (5) from an American TV broadcasting emphasis employs the segmental material of the numeral *fourteen*. The emphasis in the topic *fourteen years* brings the meaning 'fourteen years is a huge span of time'. The numeral *fourteen* has two tonal peaks — the first one is positioned on the syllable *four-* , the second one — on the syllable *-teen*. The frequency tracing demonstrates a significant rise on *four-* and a fall — on *-teen*. There is a substantial pause between *four-* and *-teen* to serve as a boundary in dividing the emphatic articulation into two phases. The waveform clearly demonstrates the break of articulation between *four-* and *-teen*. On the frequency tracing the word *fourteen* is delimited by vertical tracking.

(5) *He was on death row for fourteen years!*

Similarly, in example (6) taken from a French breaking news program the polysyllabic word *considérablement* 'significantly' displays the two components of an emphatic frequency contour: the fall on *con-* and the basic rise — on *–ment*. The rise on *considérablement*, as well as the rise on the initial topic *température*, designates the topic of sentence (5) referring to the focus *chuté ces derniers jours dans plusieurs pays d'Afrique…* with a focal fall on *chuté*. Therefore, sentence (6) has two topics. The sentence-final word *Afrique* in *chuté ces derniers jours dans plusieurs pays d'Afrique…* also carries a rise which serves here as a cue for discourse continuation, i. e. it shows that the sentence (6) is not discourse-final. On the frequency tracing the emphasized word *considérablement* is delimited by vertical tracking.

(6) *Température a considérablement chuté ces derniers jours dans plusieurs pays d'Afrique…*
'The temperature has severely fallen during the last days in many African countries…'

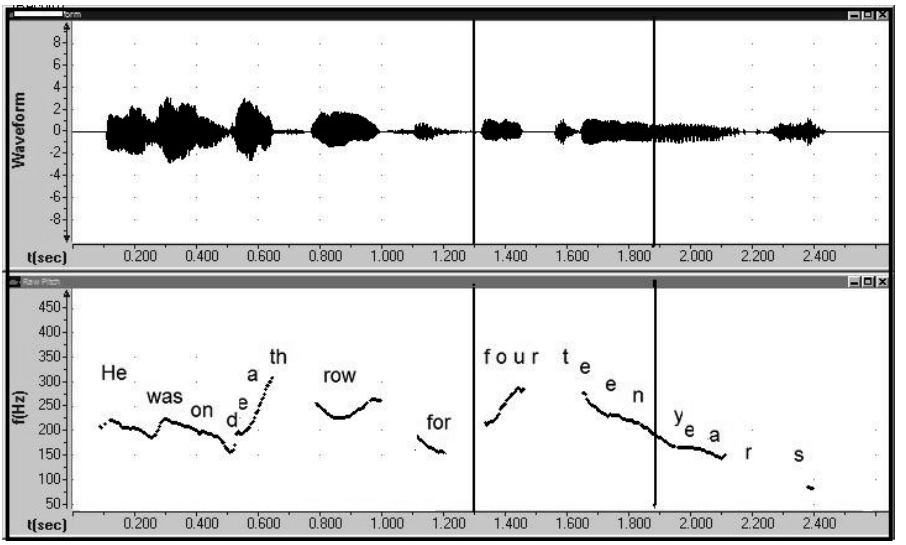


Fig. 5. The waveform and frequency tracing of Example (5)

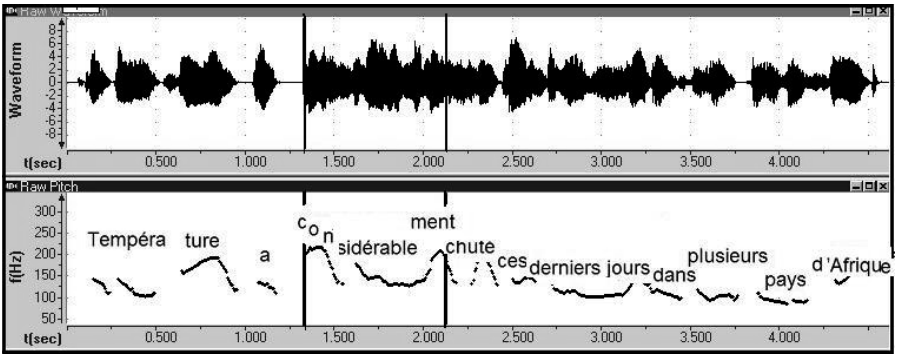


Fig. 6. The waveform and frequency tracing of Example (6)

As for the Russian language, it might be hypothesized that the so called Fifth intonational construction IK-5 as defined by E. A. Bryzgunova [Russkaia grammatika 1982: 116–117] which is characteristic of the Russian exclamations and whose frequency contour is specified by the two — the sentence-initial and the sentence-final — tonal peaks is, in fact, an instance of emphatic focus which has two accent-bearers: at the onset and at the end of a sentence. An emphatic focus can be displayed by segmental material consisting of more than one word with a sentence-initial word carrying a rise and a sentence-final word carrying a fall, and by a segmental material of only one word as well, even a monosyllabic one: *Kakaia udacha!* '*What a good chance!*'; *Skol'ko devushek!* 'How many girls!'; *Zamechatel'no*! 'Great!'; *Otvratitel'no*! 'How disgusting!'; *Styd!* 'What a shame!'. Cf. laboratory example (7) which demonstrates the parameters of Bryzgunova's IK-5:

(7) *Kakoi* *nomer*!; *Zamechatel'no*!
'What a show!' 'Great!'

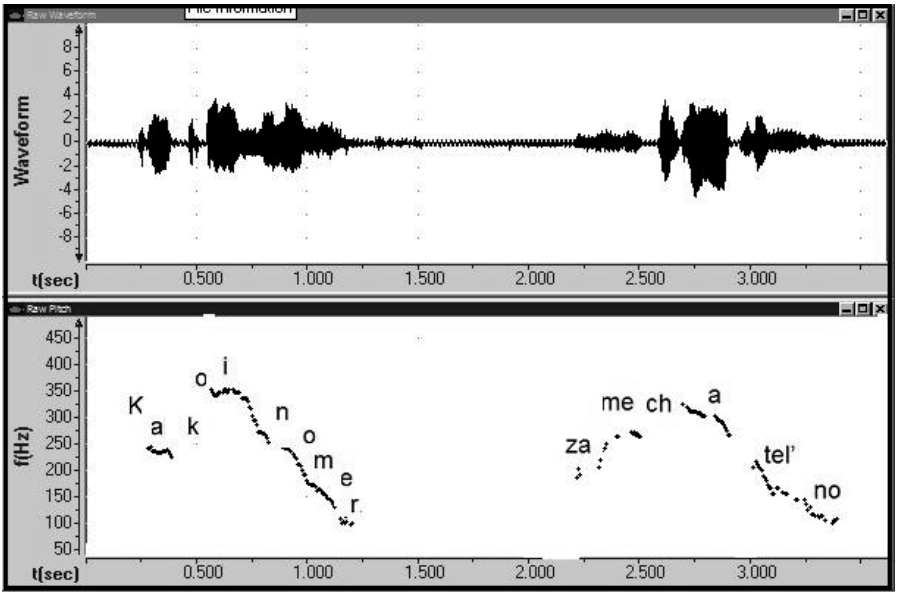


Fig. 7. The waveform and frequency tracing of Example (7)

In *Kakoi nomer*! 'What a show!' the stressed syllable [koj] in *kakoi* 'what a..!' carries a rise, while the stressed syllable [no] in *nomer* 'show, performance' carries a fall. *Zamechatel'no*! 'Great!' has two gradual rises of fundamental frequency on the initial — generally unstressed — syllables [za] and [mi] and a fall — on the stressed syllable [tʃa]. It could, therefore, be hypothesized that an emphatic focus with a monosyllabic segmental material, e. g. *Syn!* 'Son!' or *Styd!* 'Shame!', might be presented as a result of IK-5 compressing. For an emphatic topic or a question, e. g. *Syn?* 'Son?', which are based on the rise of frequency within the stressed syllable the pattern will be obviously different.

To conclude, from the viewpoint of semantic structure emphasis always functions in composition with the communicative parameters producing the speech acts of a statement or a question. Emphasis generates emphatic topics, emphatic foci, and emphatic questions. From the viewpoint of expression emphasis is specified by the following parameters:

- a longer overall duration of the stressed syllable in the accent-bearer of an emphatic communicative component as contrasted to the respective non-emphatic one;
- a longer duration of the accent-bearer initial consonants (if any) in comparison with the respective non-emphatic ones;
- a specific "emphatic tonal curving" when the basic tone — the falling one which originally designates a focus or the rising one which designates a topic

or a *yes-no*-question — associates with a tone which precedes the basic one and points in the opposite direction;
- a syllable-initial suppression of amplitude which is followed by a jump of oscillation amplitude in an emphatic communicative component.

With the only accent-bearer of an emphatic component emphasis is specified by a heterogeneous sounding with a suppressed onset and a full quality ending of the stressed syllable. In addition, emphasis can also be displayed on various segmental material and have not only one but also two accent-bearers with two stressed syllables. With two stressed syllables emphasis is displayed with the basic — depending on the communicative function either falling or rising — frequency change course on the second stressed syllable and a frequency change course pointing in the opposite direction on the first stressed syllable. Emphasis, thus, is characterized by dividing the emphatic articulatory data into two phases and lengthening both of them as being contrasted to the respective non-emphatic topics, foci, and questions. These parameters are similar in all examined languages. The set of the distinguished prosodic parameters can serve for formalizing the process of disclosing emphasis in spoken discourse.

References

1. *Faustova N. A.* Emfaza i ee intonatsionnoe vyrazhenie [Emphasis and its intonational expression]. Voprosy filologii — Problems of philology, 2006, no. 6, pp. 255–259.
2. *Féry C.* The phonology of Focus in French. Audiatur Vox Sapientiae. A Festschrift for Arnim von Stechow. Berlin, 2001.
3. *Pal'ko M. L.* Emfaza v nemeckom iazyke: dialogicheskie strategii v monologicheskoi rechi [German emphasis: dialogic strategies in monologs]. Logicheskii analiz iazyka. Mono-, dia-, polilog v raznyh iazykah i kul'turah — Logical analysis of language. Monologue and dialogue in different languages and cultures. Moscow. Indrik Publ. 2010. 295–304.
4. *Yanko T. E.* Kommunikativnye strategii russkoi rechi [Communicative strategies of the Russian speech] Moscow. Yazyki slavianskoi kul'tury Publ. 2001. 384 pp.

СИСТЕМА СЕМАНТИЧЕСКОЙ РАЗМЕТКИ КОРПУСА ТЕКСТОВ В ОГРАНИЧЕННОЙ ПРЕДМЕТНОЙ ОБЛАСТИ[1]

Загорулько М. Ю. (zagulko_maxim@yahoo.com),
Кононенко И. С. (irina_k@cn.ru),
Сидорова Е. А. (lena@iis.nsk.su)

Институт систем информатики им. А. П. Ершова СО РАН, Новосибирск, Россия

Рассматривается технология объектно-ориентированной экспертной разметки корпуса текстов, предназначенная для извлечения знаний при построении информационных систем для конкретных предметных областей. Исследуются методы и программные средства объектно-ориентированного аннотирования корпусов текстов с целью выявления терминологии и способов представления универсальных ситуаций и отношений. Предложены общие принципы терминологической разметки и разметки универсальных ситуаций и отношений, которые легли в основу разметки коллекции текстов по катализу. Представлена разработанная система экспертной семантической разметки текстов, описаны её пользовательский интерфейс, функционал и архитектура. Описаны перспективные направления использования разметки корпуса: терминологическое наполнение предметных словарей на основе терминологически размеченных фрагментов текста, создание семантико-синтаксических моделей для извлечения фактов из текста.

Ключевые слова: разметка корпусов, семантическая разметка, объектно-ориентированное аннотирование, специализированный корпус текстов

Введение

При создании лингвистических ресурсов для использования в информационных системах, ориентированных на конкретную область знаний, необходима инструментальная среда исследования корпуса текстов [3], применимая для работы экспертов. Такая среда позволила бы эксперту сопоставлять фрагменты текста заданным понятиям или категориям в соответствии с моделью предметной области (ПО). В отличие от лингвистической разметки корпуса (морфологической, синтаксической и т. п.), используемой многими исследователями [1], семантическая разметка специализированного корпуса

[1] Работа выполняется при финансовой поддержке Президиума РАН (Интеграционный проект СО РАН № 15/10 «Математические и методологические аспекты интеллектуальных информационных систем»).

ориентирована на конкретную предметную область, и должна производиться аннотаторами в соответствии с предварительно разработанными и согласованными с экспертами принципами разметки [8].

На сегодняшний день основным форматом представления семантической разметки является текст с тегами (xml, rdf, wiki и т. п.), помечающими начало и конец выделяемых фрагментов, и атрибутами, описывающими признаки фрагмента [2]. Несмотря на несомненные достоинства, связанные с развитием стандартов разметки тэгами и наличием средств их визуализации и обработки, данный подход имеет ряд недостатков, таких как:

- сложность выделения разрывного фрагмента,
- сложность описания связей между фрагментами,
- невозможность выделить описание многоатрибутных объектов предметной области (проекцию объекта на текст),
- система признаков, используемая аннотатором, «размыта» по тэгам и атрибутам,
- неэффективность программной обработки по сравнению со специализированными форматами и т. д.

Альтернативой является использование «внешнего» аннотирования, синхронизированного с текстом [7]. В этом случае описание фрагмента создается отдельно от текста и связывается с текстом указанием позиций его начала и конца. Развивая данный подход, можно создавать описания сущностей (сколь угодно сложные) и связывать их с текстом, указывая позиции начала и конца фрагментов, сопоставляемых с той или иной частью структуры описываемого объекта. Такой подход позволяет осуществлять *объектно-ориентированную разметку* текста и в значительной степени расширяет возможности использования разметки корпуса для создания лингвистических ресурсов, ориентированных на анализ текстов предметной области (терминологических словарей, шаблонов и правил анализа).

В данной работе предложен подход к семантической разметке текста, позволяющий сопоставлять объектно-ориентированные представления сущностей фрагментам текста. Приводимые иллюстрации основаны на результатах проекта по созданию специализированного семантически размеченного корпуса текстов по катализу [4].

1. Принципы семантической разметки

При построении информационных систем неизбежно возникает задача автоматизации процесса извлечения экспертных знаний о предметной области и ее подъязыке — системе понятий и отношений между ними, способах представления сущностей и типовых ситуаций предметной области. Такая задача эффективно решается методами корпусной лингвистики, то есть путем создания и исследования специализированного корпуса текстов, представляющего собой достаточный объем снабженных экспертной интерпретацией

лингвистических данных, который может служить основой формирования системы автоматического анализа текстов, т. е. выступать в роли обучающего корпуса. В состав корпуса текстов отбираются фрагменты из справочной и учебной литературы, научные статьи и рефераты, посвященные определенной тематике.

При создании специализированных корпусов текстов обычно производится лингвистическое аннотирование (морфологическое, синтаксическое), не зависящее от ПО и осуществляемое автоматически и/или вручную. Семантическая разметка, напротив, предметно ориентированна, поскольку определяется онтологией ПО и производится экспертами [8]. Процессу семантической разметки специализированного корпуса текстов предшествует достаточно длительный (2–3 месяца) предварительный этап совместной работы экспертов, лингвистов и разработчиков системы, в рамках которого происходит обмен компетенциями, выработка и согласование признаков и принципов разметки. Результатом этого этапа является инструкция по семантическому аннотированию. Речь идет о двух видах семантического аннотирования:

- терминологическая разметка, которая в первую очередь предназначена для фиксации в тексте имен понятий ПО,
- разметка отношений (или ситуаций, представляющих собой многоместные отношения), в которых размеченные сущности выступают в определенных семантических ролях.

Ниже изложены основные принципы аннотирования корпуса текстов по катализу и результаты предварительных экспериментов по разметке.

Терминологическая разметка фиксирует не только присутствие в тексте наименований сущностей ПО, но и особенности использования общеупотребительной лексики в данном подъязыке. Предложены следующие принципы терминологической разметки:

- Определять признаки, соответствующие типам и подтипам сущностей в соответствии с иерархией признаков.
- Определять максимальный текстовый фрагмент, представляющий сущность. *(нитрил акриловой кислоты)*. При этом остается возможность указывать и вложенные фрагменты, представляющие сущности (*нитрил акриловой кислоты*).
- Сопоставлять аннотацию самой сущности, а не всей синтаксической группе, в которую она входит *(в реакцию сочетания вступают кислородсодержащие продукты CH3OH, CH2O)*.
- Размечать все ссылки на сущности, в том числе анафорические замены.
- Использовать при необходимости разрывные фрагменты (*после контакта CH4 с оксидами ряда металлов*), в том числе при перечислении с сочинительным сокращением: *конденсация гликолевого и глицеринового альдегидов*

Иерархия признаков для терминологической разметки текстов по катализу позволяет пометить фрагмент текста как Вещество или конкретный подкласс веществ (Элемент, Соединение и др.). В иерархию признаков для веществ

внесены и лексические показатели Роли (ролевая лексика *катализатор, реагент, реакционная смесь, реактант, продукт* и т. п.).

Кроме того, система признаков представляет классы предикатов:

- Реакции:
  - Химические реакции (*окисление, гидрировать, крекинг)*, подклассы которых учитывают валентностный потенциал: именные реакции (*реакция Будуара*), реакции с инкорпорированным участником (*метанирование*) и др.;
  - Обобщенные реакции (Взаимодействие, Превращение, Получение);
- Лексические показатели ситуаций (в частности, представляющие взаимосвязи процессов и веществ в составе ситуации (такие как *катализировать*, *приводить к, использоваться в качестве/для*).

Разметка отношений (ситуаций) производится над терминологически размеченным текстом. Рассмотрим ситуации типа ПРОЦЕСС, описывающие процессы молекулярного взаимодействия в катализе:

Паровая конверсия метана в синтез-газ протекает на никелевом катализаторе.

Такая микроситуация представляется как многоместное отношение:

ПРОЦЕСС (Реакция, Реагент, Катализатор, Продукт)

где Реакция (*паровая конверсия*) — химическая реакция, характеризующаяся превращением одного или нескольких исходных веществ, выступающих в семантической роли Реагентов (*метан*), в отличающиеся от них по химическому составу или строению вещества, Продукты реакции (*синтез-газ*), с участием Катализатора (*никель*).

Ниже приводится пример разметки текста, в котором каждое вхождение термина и ситуации выделено идентификатором (Ti, Pj). Из примера видно, что *этилен* выступает в роли Продукта в ситуации П44 и в роли Реагента — в ситуации П46, таким образом, фрагменты, описывающие ситуации, в данном случае пересекаются:

Кроме того, этилен <T55>, *образующийся* <T56> *при окислительном превращении* <T57> *CH*4 <T58>, *можно почти полностью конвертировать* <T59> *в другие олефины* <T60> *и ароматические углеводороды* <T61> *на цеолитах* <T62>.

Процесс П44 (Реакция: T56 <Образование>, Продукт: T55)

… *этилен* <T55>, *образующийся* <T56> …

Процесс П45 (Реакция: T57 <Химическая реакция>, Реагент:T58)

…*этилен* <T55>, *образующийся* <T56> *при окислительном превращени*<T57> *CH*4 <T58>,…

Процесс П46 (Реакция: T59 <Превращение>, Реагент:T55, Продукт: T60, Продукт: T61, Катализатор: T62)

…*этилен* <T55>, *образующийся* <T56> *при окислительном превращении* <T57> *CH*4 <T58>, *можно почти полностью конвертировать* <T59>

в другие олефины <Т60> *и ароматические углеводороды* <Т61> *на цеолитах* <Т62>

Разработаны следующие принципы ситуационной разметки:

- Разделение терминологической и ситуационной разметок. Так, во фразе: *В данной работе мы исследовали влияние предварительного восстановления водородом платиновых и палладиевых катализаторов*
  терминологическая разметка:
  палладиевых <Элемент, Экземпляр> *катализаторов* <Роль>
  ситуационная разметка:
  палладиевых <Реагент, П24>
- Диагностирующим контекстом, позволяющим предположить наличие в тексте описания ситуации/отношения, является присутствие в нем соответствующего лексического предиката (для химических процессов — присутствие термина-обозначения конкретной или обобщенной реакции).
- Ситуация обычно выражена в рамках клаузы/предложения, выход за рамки клаузы/предложения возможен при анафорической замене, при этом в разметке указывается антецедент:
  первой стадией является синтез метанола, далее следуют его дегидратация
- При сочинении различаются множественные ситуации и ситуации с множественными участниками.
  Глубокое окисление метана на платиновых и палладиевых катализаторах, нанесенных на нитрид кремния (2 ситуации)
  Разложение сероводорода на элементную серу и водород (1 ситуация, 2 Продукта)
- Лексические показатели ситуаций не создают новых ситуаций, но фиксируют связи в рамках ситуации, позволяя определить потенциальных участников.
  Метанол подвергается превращениям, которые характерны для катализа высококремнистыми цеолитами.

2. Архитектура системы разметки текста

Внутренняя работа системы основывается на трех концептах: дерево признаков, разметка текста и список ситуаций (см. Рис.1).

Дерево признаков реализует представление иерархической системы признаков, где нулевая вершина является фиктивным признаком, не участвующим в разметке текста, а любой другой элемент дерева может одновременно выступать как в роли признака, так и в роли вершины поддерева признаков. Для хранения разметки текста используется хеш-таблица, в которой ключу соответствует признак, по которому размечен фрагмент текста, а значению соответствует начальная и конечная позиции в тексте и список экземпляров ситуаций. Список отношений (ситуаций) содержит их абстрактные описания, на основе которых создаются конкретные экземпляры.

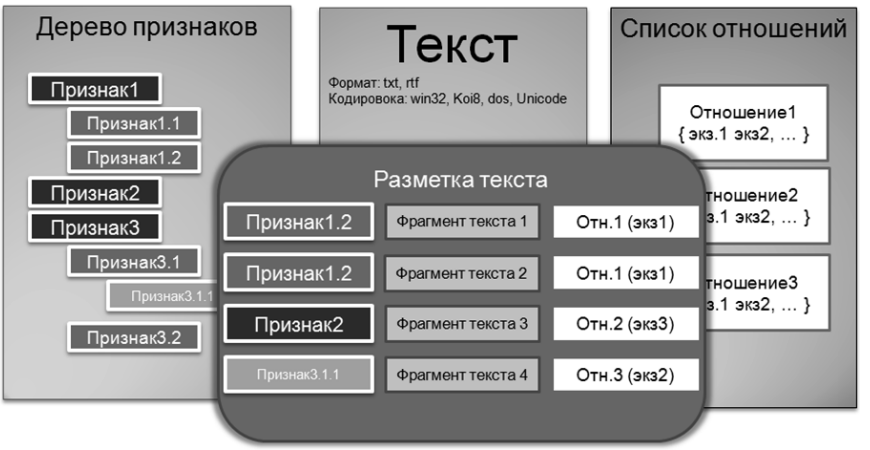


Рис. 1. Архитектура системы разметки текста

Разработанный инструмент позволяет пользователю формировать иерархию признаков, каждому признаку сопоставлять цветовую и стилевую схему разметки, которая используется при реализации функций визуализации разметки в тексте. Поддерживается раздельный просмотр разметки по признакам/ группам признаков.

Описание фрагментов вынесено в отдельную таблицу, в которой отражены позиции, текстовое представление, признаки и связи фрагментов. Таблица поддерживает навигацию в тексте, а также сортировку фрагментов по разным параметрам.

Для поддержки всех свойств разметки, а также для более эффективного дальнейшего использования для автоматизированного создания лингвистических ресурсов принято решение отказаться от стандартного формата хранения размеченного текста в виде текста с xml-тегами, помечающими начало и конец выделяемых фрагментов. Вместо этого создается аннотация, синхронизированная с исходным текстом (текст загружается из тестового или rtf-файла и в дальнейшем не меняется). Аннотация — это множество троек <признак, позиция, информация>, которые фиксируют, что определенная символьная последовательность в тексте (фрагмент) обладает определенными свойствами. В процессе работы аннотатора разметка динамически визуализируется.

Пользовательский интерфейс системы разметки текстовых корпусов, разрабатываемой для лингвистов и экспертов предметной области, должен быть легким (интуитивно понятным) как для опытного, так и для начинающего пользователя.

Система разметки предоставляет следующие возможности.

- Загрузка текста формата txt, rtf, поддержка кодировок win32, Koi8, dos, Unicode.
- Загрузка и сохранение размеченного текста в специальный формат mspr (Mars System Project).

- Просмотр и редактирование дерева иерархии признаков.
- Просмотр и редактирование фрагментов, приписанных определенному признаку.
- Просмотр всех (или части) размеченных фрагментов одновременно (в видимой части текста).
- Загрузка и сохранение размеченного текста; формат файла текстовый, например, xml-подобный.
- Сортировка списка размеченных фрагментов текста по позиции в тексте, по имени признака, по фрагменту текста.
- Обеспечение многослойной разметки.

На Рис.2 представлен пользовательский интерфейс системы разметки текста и продемонстрирована ситуационная разметка (терминологическая отфильтрована), при этом разметка ПРОЦЕССов частично перекрыта разметкой другими отношениями, например в первой фразе текста процессы *Окислительная конденсация метана* и *Синтез этана, этилена и других углеводородов* связаны отношением Реализации (отражающим способ осуществления химической реакции). Помимо ситуаций, в тексте отражены лингвистические отношения, например во фразе …*высокая стабильность CH4 затрудняет его переработку* светло-серым цветом показана анафорическая связь, где *CH4*- антецедент, а местоимение *его* — анафор.

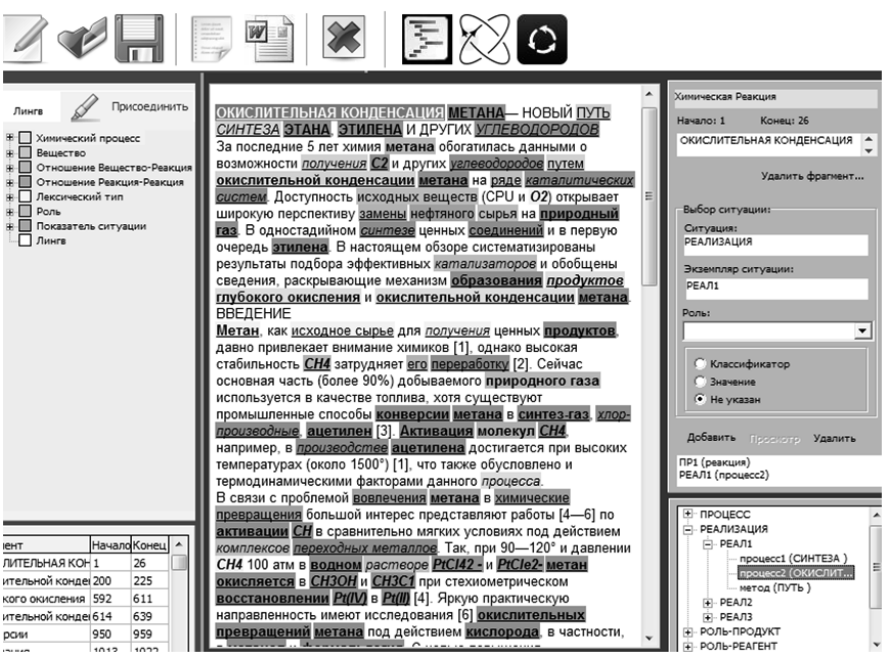


Рис. 2. Система разметки текста

3. Применение размеченного корпуса

Создание языковых ресурсов, ориентированных на автоматическую обработку текста, — довольно трудоемкий процесс, поэтому естественной является попытка автоматизировать их создание и начальное наполнение на основе аннотированного корпуса текстов (в первую очередь имеется в виду экспертная семантическая разметка).

Мы выделили следующие перспективные направления автоматизации процесса создания лингвистических ресурсов:

– Терминологическое наполнение предметных словарей;
– Создание семантико-синтаксических моделей для извлечения фактов из текста.

Первое направление достаточно изучено [5], однако с учетом наличия разрывных фрагментов и многословных терминов требует технологических пояснений. Для второго направления предложены основные идеи и принципы создания подобного ресурса.

3.1. Наполнение терминологических словарей

Наполнение предметного словаря на основе терминологически размеченных фрагментов текста осуществляется в несколько этапов (см. Рис.3).

- Перенос иерархии семантических признаков в словарь и согласование их с уже имеющимися в словаре признаками.
- Обработка текстовых фрагментов морфологическими и синтаксическими компонентами словарной технологии.
- Нормализация и формирование терминов (для многословных фрагментов фиксируется синтаксический шаблон или, если такой шаблон не найден, то фрагмент добавляется как несогласованный значимый словокомплекс).
- Снабжение терминов семантическими признаками в соответствии с размечаемыми признаками.

В ходе наполнения словаря возникают технические моменты, рассмотрение которых требует отдельного внимания:

(1) *Морфологическая и лексическая омонимия*. Поскольку эксперт не осуществляет морфологической и синтаксической разметки, то в словарь добавляются и снабжаются семантическими признаками все омонимы, соответствующие термину.

(2) *Универсальная лексика, входящая в состав словокомплексов* (многословных терминов). В этом случае словарь расширяется универсальной лексикой с пометкой о ее нетерминологичности.

(3) *Несловарная лексика, отсутствующая в универсальном словаре русского языка*. Неизвестные слова могут встречаться в качестве однословного

термина или входить в состав размеченного многословного фрагмента. В данном случае используется предсказание, которое строит гипотезу о принадлежности слова той или иной части речи, а также его морфологических и лексических признаках.

(4) *Разрывный фрагмент* рассматривается как синтаксическая группа, поэтому для него также формируется словокомплекс в соответствии с найденным синтаксическим шаблоном.

(5) *Буквенно-символьные конструкции* не являются элементами словаря, однако являются терминами в данной ПО и образуют синтаксическую и семантическую связь, входя в состав словокомплекса.

Рис. 3. Схема наполнения терминологических словарей

Наличие автоматически-пополняемого терминологического словаря, в случае, когда пополнение происходит параллельно с работой пользователя, позволяет интерактивно использовать новые термины для автоматизации дальнейшей работы эксперта. Т.е. словарный компонент может осуществлять автоматический поиск новых терминов при последующей разметке текстов. Ошибки, возникающие вследствие того, что либо фрагмент размечен неправильно, либо термин некорректно описывается в словаре, фиксируются и служат основанием для корректировки экспертом.

3.2. Семантико-синтаксические гипотезы

На основе разметки ситуаций или, в общем случае, разметки отношений можно фиксировать синтактико-семантические модели, или шаблоны (в частности, модели управления предикатных лексических единиц). Предполагается, что вначале создаются гипотезы в виде шаблонов, для которых накапливается статистика встречаемости и на основе накопленных данных впоследствии принимается решение о достоверности шаблона. Далее можно применять дополнительные методы обобщения шаблонов или непосредственно предоставлять результаты пользователю для обработки.

В качестве примера такого отношения и шаблонов его распознавания в тексте можно рассмотреть следующий:

Литература

1. *Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л.* и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003–2005. –М.:Индрик, 2005. — С.193–214.
2. *Биряльцев Е. В., Елизаров А. М., Жильцов Н. Г., Иванов В. В., Невзорова О. А., Соловьев В. Д.* Модель семантического поиска в коллекциях математических документов на основе онтологий // Труды 12й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2010. –Казань, 2010. — С.296–300.
3. *Захаров В. П.* Корпусная лингвистика: Учебно-метод. пособие. — СПб, 2005. — 48с.
4. *Кононенко И. С., Сидорова Е. А.* Система семантической разметки корпуса текстов как инструмент извлечения экспертных знаний (на материале текстов по катализу) // Труды международной конференции «Корпусная лингвистика — 2011». — Санкт-Петербург, 2011. — С. 193–198.
5. *Ляшевская О. Н., Сичинава Д. В., Кобрицов Б. П.* Автоматизация построения словаря на материале массива несловарных словоформ // Браславский П. И. (отв. ред.), Интернет-математика — 2007: сборник работ участников конкурса научных проектов по информационному поиску. –Екатеринбург: Издательство Уральского университета, 2007. –С.118–125.
6. *Яковчук Е. И., Сидорова Е. А.* Обобщенные семантико-синтаксические модели в задачах обработки текста // Труды рабочего семинара «Наукоемкое программное обеспечение НПО-2011». Ершовская конференция по информатике. –Новосибирск: ИСИ СО РАН, 2011. –С.287–292.
7. *Blanco X.* Using NooJ for Multipurpose Analysis of Romance Languages Corpora // Труды международной конференции «Корпусная лингвистика–2008». – СПб., 2008. — С.40–44.
8. *Kim J. D., Ohta T., Tsujii J.* Corpus annotation for mining biomedical events from literature // BMC Bioinformatics. 2008. 9:10

РУССКОЕ *РАЗОЧАРОВАНИЕ* В ЕВРОПЕЙСКОМ КОНТЕКСТЕ: ИСТОРИЯ И СОВРЕМЕННОСТЬ[1]

Зализняк А. А. (anna.zalizniak@gmail.com)
Институт языкознания Российской Академии наук

Шмелев А. Д. (shmelev.alexei@gmail.com)
Московский педагогический государственный университет

В докладе рассматривается семантическая история слова разочарование (а также разочаровать, разочароваться, разочарован) в русском языке. Показано, что слово разочарование претерпело в русском языке определенную семантическую эволюцию, в результате которой оно, сохранив, в значительной степени, исходное значение франц. désenchantement, существенно разошлось с современным «общеевропейским» концептом, соответствующим значению слов англ. disappointement, франц. déception, нем. Enttäuschment и др., обозначающих чувство обманутого ожидания, вызванное оценкой некоторого внешнего по отношению к субъекту события. Между тем русское разочарование описывает событие внутреннего мира: состояние, вызванное утратой ценности, которое видоизменяет индивидуальную картину мира. В современном русском языке, в соответствии с общей тенденцией семантического калькирования с английского, значение слова разочарование и его производных выравнивается под значение англ. disappointment. В докладе предлагается семантическое описание для стандартного значения обсуждаемых русских слов и демонстрируется, в чем именно состоит происходящий на наших глазах, но при этом плохо идентифицируемый носителями языка, сдвиг в их значении и употреблении.

Ключевые слова: семантика, семантическая эволюция, калька, русский язык, французский язык, английский язык, внутреннее состояние, ментальное состояние, эмоции, оценка

В отличие от имен многих других внутренних состояний (таких как *страх*, *гнев*, *радость*, *счастье*, *сожаление*, *сомнение*, *подозрение*, *тоска*), которым посвящена целая серия монографических описаний, слово *разочарование*, будучи обозначением не менее важного чувства, похоже, вообще не проанализировано: соответствующей статьи нет в Толково-комбинаторном словаре [Мельчук, Жолковский 1984], в книге [Виноградов 1994], в словаре [Черных 1994], и нам неизвестно ни одно специальное лингвистическое исследование

[1] Статья написана при финансовой поддержке РГНФ (в рамках научно-исследовательского проекта РГНФ «Эволюция русского лексикона в европейской лингвистической перспективе», проект № 11-04-00105а).

на эту тему; мы располагаем лишь толкованием, предложенным в известной работе [Иорданская 1970]. Между тем существительное *разочарование*, также как и глаголы *разочаровать* и *разочароваться*, не говоря о культурной значимости заключенного в них концепта, представляют исключительный интерес в разных отношениях.

В значении предикатов, описывающих эмоции, обычно выделяют два основных компонента, характеризующих установку субъекта по отношению к ситуации-объекту: оценка этой ситуации по шкале «хорошо — плохо» и оценка ее по признаку существования/вероятности — см. Wierzbicka 1969: 108; Иорданская 1970, Wierzbicka 1972: 60–64; Зализняк 1983; Searle 1983; Зализняк 2006: 74–86; 478–487, Goddard 1997: 89; ср. также компоненты “something good happened” и “something bad happened” в толковании имен эмоций в Wierzbicka 1999. (Эта традиция восходит к исследованиям природы эмоциональных состояний в трудах Лейбница, Спинозы и Декарта.) При этом многие «имена чувств» имеют значения, где центральным является оценочный компонент и такие, где главной оказывается оценка вероятности (ср. Зализняк 1983; Iordanskaja, Mel’čuk 1990; Зализняк 2006 — о глаголах *бояться*, *надеяться* и *ждать*). Как мы увидим дальше, в значении слов, описывающих разочарование, присутствует та же вариативность: имеется класс употреблений, который ставит в центр экзистенциальный компонент (‘X считал, что нечто хорошее произойдет/имеет место, а оно не произошло/не имеет места’), и другой, где центральным оказывается оценочный (‘раньше X оценивал нечто хорошо, а теперь — плохо’).

Начнем с того, что слово *разочарование* является прозрачной морфологической калькой с французского *désenchantement*, и, очевидно, изначально представляет собой результат импорта важнейшего европейского культурного концепта эпохи романтизма. По данным Национального корпуса русского языка слова *разочарование*, *разочарованный*, *разочаровать* появляются в начале XIX века, причем до 1830 это буквально считанные примеры.

Далее, на протяжении XIX-XX вв. сформировался современный «общеевропейский» концепт, соответствующий значению слов англ. *disappointment*, франц. *déception*, нем. *Enttäuschung* и др., обозначающих своего рода чувство обманутого ожидания. Что же касается русского слова *разочарование*, то оно в большей степени сохранило первоначальное значение, отсылающее, в соответствии с его внутренней формой, к идее предшествующего «очарования» и выражаемое наиболее явно в конструкции *разочарование, разочаровался, разочарован* ***в ком/в чем***. Однако, поскольку граница между упомянутыми вариантами значения трудноуловима (наша задача состоит именно в том, чтобы ее выявить), русское слово *разочарование* взяло на себя функцию переводного эквивалента для европейских слов англ. *disappointment*, франц. *déception*, нем. *Enttäuschung* и др., что отражается как в соответствующих двуязычных словарях, так и в практике перевода. В каких-то случаях такая эквивалентность действительно имеет место, однако в других — вызывает сомнения (ниже будут рассмотрены случаи последнего типа).

Завершающий этап этой истории состоит в том, что сегодня русское слово *разочарование*, и в особенности глагол *разочаровать*, стали утрачивать свою

специфичность, т. е. приобрели возможность употребления, сходного с англ. *disappointment, to disappoint*. Как в звучащей речи вокруг нас, так и в Интернете, нам все чаще приходится сталкиваться с такими употреблениями, которые выходят за грань прежней нормы и вызывают подозрение в том, что являются калькированными.

Прежде чем перейти к семантическому анализу, приведем примеры употребления глагола *разочаровать*, соответствующего старой норме и отклоняющегося от нее. Носителем «общеевропейского концепта» разочарования par excellence мы будем считать англ. глагол *to disappoint* (мы также будем обращаться к близким по значению, хотя и, естественно, не полностью совпадающим, словам нем. *enttäuschen* и франц. *decevoir*).

То, что мы называем старой нормой, или «образцовым» разочарованием, может быть проиллюстрировано следующими примерами:

(1) *Глебов был **разочарован**: он считал пса грозным и смелым, а тот вёл себя как трусишка.* [Юрий Трифонов. Дом на набережной (1976)][2]

(2) *Знаете ли, Владлен, я тоже очень любил Смоктуновского в «Идиоте». Но посмотрел «Царя Федора» и **разочаровался в нем**.* [Владлен Давыдов. Театр моей мечты (2004)]

(3) *…лучше было на ялтинском берегу воображать неаполитанскую ночь — тем более что неаполитанская при ближайшем рассмотрении могла оттолкнуть и разочаровать, как **разочаровал** его когда-то Париж.* [Дмитрий Быков. Орфография (2002)]

Употребление, отклоняющееся от старой нормы, может быть проиллюстрировано следующими примерами. Их по крайней мере два класса: (I) употребление в контексте событийного имени или придаточного с *что* и (II) употребление глагола в несов. виде в стативном значении.

I. Употребление в контексте событийного имени:

(4) *Он был **разочарован ее отказом***

(5) *Она была **разочарована его женитьбой***

(6) *Он был **разочарован рождением дочери***

(7) *За выходные он убедил себя, что предстоит женитьба, и теперь был почти **разочарован** ее отказом (почти разочарован ≅ почти огорчен)*

[2] Здесь и далее примеры со ссылками, как правило, взяты из Национального корпуса русского языка, прочие — по большей части найдены в Интернете при помощи поисковых систем Яндекс и Google.

(8) *Он был так* ***разочарован*** *ее отказом надеть красивое платье, что она в конце концов смилостивилась (так разочарован ≅ так сильно огорчен)*

(9) *Похоже, он не слишком* ***разочарован*** *ее замужеством!*

Ср. примеры аналогичного употребления англ. *disappointed* :

(10) *I said I was* ***disappointed by his refusal****. He said, 'You can't always get what you want.'*

(11) *Eager to show the world that she was not* ***disappointed by his departure****, and apparently in a pique, she accepted a proposal of marriage from John Marshall …*

(12) *Fully expecting to receive an offer of marriage from Onassis, she was then bitterly* ***disappointed by his marriage*** *to Jacqueline Kennedy*

Далее (конструкция с изъяснительным придаточным):

(13) *Игра была отличной, но я* ***очень разочарован, что*** *не был засчитан первый гол.*

(14) *Ей-богу, он был* ***разочарован, что*** *приехал я, а не вы… [Вера Белоусова. Жил на свете рыцарь бедный (2000)]*

(15) *Эрвин был завидным женихом, и многие женщины были* ***разочарованы, что*** *место рядом с ним оказалось занято мной.*

Ср. примеры аналогичного употребления англ. *disappointed that*:

(16) *Spanish official* ***disappointed that*** *Obama will not attend EU-US summit.*

(17) *Foreign Trade Minister Fientje Moerman said that she was* ***disappointed that*** *the Conference, which "had so many hopes riding on it, ended in failure for everyone".*

II. Употребление глагола в несов. виде в стативном значении:

(18) *Почему ты так коротко отвечаешь на мои письма?* ***Это разочаровывает*** *и огорчает.*

(19) *А когда вы ведете себя как маменькины сыночки — это* ***разочаровывает…*** *и вся любовь может перерасти в презрение….*

(20) *Вопрос: Что вас* ***разочаровывает*** *в мужчинах? Ответы: Ложь; Холодность, отчуждение, равнодушие…; Педики! Меня очень* ***разочаровывает****,* ***что*** *они среди нас)))*

(21) *Светлана Кузнецова: «**Разочаровывает, что** не смогла показать свой теннис»*

(22) *Второе, что **разочаровывает, что** изображение, видимое в затворных очках, заметно темнее…*

Ср. примеры аналогичного употребления англ. *it disappoints/ it is disappointing that*:

(23) ***It disappoints** me greatly that the programme is completely uncritical of the practices of supermarkets*

(24) *TCMP player released in new version: 0.70 — but **it disappoints** very much.*

(25) ***It is disappointing** that we seem to observe the world as through a membrane*

(26) ***It is disappointing** that Rob's e-mail is based more on emotion, hearsay and incorrect facts*

Оба эти класса употреблений русского глагола *разочаров(ыв)ать* представляются ненормативными. Прежде чем переходить к вопросу о том, что здесь нарушено, приведем примеры хорошей межъязыковой эквивалентности для употреблений, находящихся в рамках строгой нормы.

(27) *Ben, as he scanned the buildings on the Rapenburg Canal, was somewhat **disappointed** in the appearance of the great University of Leyden.* [Mary Mapes Dodge. Hans Brinker or the Silver Skates]

*— Осматривая здания на Раненбургском канале, Бен был слегка **разочарован** внешним видом славного Лейденского университета.*

(28) *Die andern lachten mich aus, weil ich mir die Berge höher vorgestellt hatte. Überhaupt bin ich fast immer **enttäuscht**, wenn ich etwas sehe, das ich mir irgendwie vorgestellt habe. Es ist nie so überwältigend, wie ich es haben wollte.* (www.dwds.de)

*Другие смеялись надо мной, потому что я представлял себе горы более высокими. Вообще я почти всегда бываю **разочарован**, когда я впервые вижу что-то, что я представлял себе заранее. Это всегда оказывается менее впечатляющим, чем мне бы хотелось.*

(29) *Dans le centre ville, un écran géant diffuse les cérémonie dès 9h du matin, pour des grand-mères venues avec le thermos et des petites filles déguisées en princesses. Venue en badaud, Fiona commente : "Je suis **désenchantée** par cette famille depuis ce qu'elle a fait à Diana. Mais William, c'est différent… ".*

[…] Я ***разочаровалась*** *в этой семье после того что они сделали с Дианой, но Уильям, это другое…*

(30) *Mais la cause de tout cela, c'est que je suis* ***désenchantée*** *par un peu tout, j'analyse, je regarde autour de moi, je fais le bilan de ma vie et rien* **...**

Но причина в том, что я ***разочаровалась*** *во всем, я все анализирую, смотрю вокруг себя, подвожу итоги своей жизни, и ничего…*

Однако в каких-то других случаях эквивалентность оказывается неполной (и, соответственно, русская фраза звучит несколько фальшиво). Пример из (вообще говоря, очень хорошего) перевода Н. Волжиной «Саги о Форсайтах» Дж. Голсуорси:

(31) *But Irene had not moved; in her gloved hands she ceaselessly turned and twisted the little bunch of violets.*
The tears of rage and ***disappointment*** *rolled down June's cheeks.*
"How could you come?" she said. "You have been a false friend to me!"
Again Irene laughed. June saw that she had played a wrong card, and broke down.
"Why have you come?" she sobbed. "You've ruined my life, and now you want to ruin his!" [John Galsworthy. The Man of Property]

Но Ирэн не двигалась; её руки, затянутые в перчатки, вертели и теребили букетик фиалок.
Слезы ярости и ***разочарования*** *заливали щеки Джун.*
— Как вы могли прийти! — сказала она. — Хороший же вы друг!
Ирэн опять засмеялась. Джун поняла, что сделала неправильный ход, и громко заплакала.
— Зачем вы пришли? — промолвила она сквозь рыдания. — Вы разбили мою жизнь и ему хотите разбить!

Чувство героини, которая узнала, что ее жених любит другую женщину, обозначенное англ. словом *disappointment*, по-русски никак нельзя назвать *разочарование*. Удивительным образом этому чувству в русском языке нет названия.

Джун могла бы *разочароваться в своем женихе*, но это совсем не то, что здесь имеется в виду. Англ. *disappointment* здесь обозначает чувство, вызванное тем, что не произошло (и понятно, что не произойдет) то хорошее, на что она рассчитывала: замужество с любимым и любящим ее человеком. Это хорошо соответствует толкованию *disappointment*, сформулированному Вежбицкой: *disappointment* — чувство, рациональная основа которого состоит в том, что нечто хорошее, что должно было произойти и на что человек рассчитывал, не произошло[3]:

[3] Ср. также [SOED: 691] " break off what has been appointed; fail to keep an engagement". Т. е внутренняя форма англ. *to disappoint* это что-то вроде «отмена запланированного».

[Wierzbicka 1972: 64]:

> “X feels as one does when one thinks that what one has desired to happen and thought would happen has not happen”

[Wierbicka 1999:71]:

> I thought that something good would happen;
> I felt something good because of this;
> I know now: this good thing will not happen.

Интересно, что толкование Иорданской очень похоже на толкование Вежбицкой, и оно тоже на самом деле больше подходит к *disappointment*, чем к русскому *разочарованию* ([Иорданская 1970: 25]; приводим с небольшими сокращениями):

> ***А испытывает разочарование*** [по поводу В] = А испытывает пассивно-отрицательное эмоциональное состояние, каузированное тем, что:
> 1) А уверен в неосуществлении события В;
> 2) до t_o А был уверен в осуществлении события В;
> 3) событие В очень желательно для А.

Например:

> Он надеялся встретить ее там, и узнав, что она не придет, ***испытал*** острое ***разочарование***.

Хотя приведенный Л. Н. Иорданской пример под это толкование вполне подходит, но это очень периферийное употребление слова *разочарование* (близкое к *disappointment*). Поэтому, в частности, слово *разочарование* не подходит для описания чувства Джун в рассмотренном выше примере, при том что оно соответствует толкованию Иорданской.

Итак, в чем же дело? Очевидно, в том, что помимо перечисленных компонентов (‘считал, что нечто произойдет’; ‘хотел, чтобы произошло’; ‘это не произошло’ — или, точнее — ‘вместо этого произошло нечто другое, худшее’), русское *разочарование* включает еще некоторый дополнительный смысл.

А именно, следующий. «Образцовое» разочарование — это такое, когда происходит утрата некоторой ценности, в результате чего заменяется фрагмент картины мира (ср. *разочарование в людях, в женщинах, в любви, в идеалах*). Этот концепт, во-первых, в русском языке оформлен особой конструкцией (*разочарование в ком/чем*), а во-вторых, и это самое главное, это исходное значение как бы отбрасывает тень на остальные (это явление Д. Н. Шмелев называл эпидигматическими связями), т. е. влияет на все типы употребления, в том числе на те, где имеется в виду непосредственная эмоциональная реакция на послужившее стимулом единичное событие. Для русского *разочарования* произошедшее (или не произошедшее) событие всегда является не объектом, а лишь основанием оценки; это лишь «явление», которое позволяет обнаружить некоторую стоящую за ним «сущность»: событие оценивается (как плохое) не само по себе, а лишь как свидетельство несовершенства мира (в том или

ином его фрагменте). И именно поэтому слово *разочарование* не уместно в случае Джун Форсайт: факт, который стал ей известен, означал для нее огромную жизненную утрату — но не утрату ценности в картине мира.

В русском языке отрицательная эмоция, возникшая из-за того, что произошло не то событие, которого человек ожидал и хотел, а нечто другое, худшее, не ограничивается этим локальным чувством, а сопровождается некоторым изменением картины мира (и поэтому здесь даже нельзя провести четкой границы между кратковременным разочарованием-реакцией и долговременным разочарованием-отношением, выражаемым конструкцией с предлогом *в*; возможно, здесь играет роль живость внутренней формы: *разочарование* так или иначе связывается с *очарованием*[4]). Если этот момент полностью отсутствует, то разочарование с точки зрения русского языка выглядит неполноценным (тогда это не разочарование, а огорчение или сожаление, т. е. специфическая семантика пропадает, остается лишь более общая). И действительно, в тех примерах, где указанный компонент отсутствует, *разочарован* близко по смыслу к *сожалеет*, *огорчен* или *расстроен*, ср.:

(32) *Шумахер: Я очень* ***разочарован, что*** *не смог помочь Ferrari*

(33) *Госдеп США* ***разочарован, что*** *убийцы американского журналиста Пола Хлебникова не привлечены к ответственности*

(34) *Главный тренер сборной России* ***сожалеет, что*** *в полуфинале Евро-2008 россияне не сыграют с Италией.*

(35) *Вторая ракетка мира Динара Сафина* ***сожалеет, что*** *американки Серена и Винус Уильямс не принимают участия в турнире в Индиан-Уэллсе*

(36) *Блаттер:* ***огорчен, что*** *России не будет в ЮАР*

(37) *Председатель совета директоров BP Питер Сазерленд* ***огорчен, что*** *руководство России не препятствует «олигархам» применять «силовые тактики»*

(38) *Я немного* ***расстроен, что*** *не смог победить нокаутом, но доволен, что удалось выиграть все 12 раундов*

(39) *Николай Давыденко* ***расстроен, что*** *стартовый матч на «Кубке Кремля» ему придется провести против соотечественника Марата Сафина*

Англ. *disappointment* ничего подобного не содержит: это противоречие ожиданиям, даже не обязательно положительным; это может быть, в соответствии

[4] Не случайно многие русские авторы в один голос говорят, что не может разочароваться, тот, кто не был очарован.

с этимологией, строго «нарушение назначенного» и ничего более, ср. англ. перевод следующего предложения из «Исповеди сына века» Альфреда де Мюссе:

(40) *Je m'attendais qu'elle allait insister pour notre projet d'aller à Genève, et en effet* ***elle n'y manqua pas*** *(A. de Musset. La Confession d'un enfant du siècle)*

(41) *I was expecting her to insist that we carry out our plan of going to Geneva, and was* ***not disappointed***.

Тем самым различие между русским концептом разочарования и современным «общеевропейским» концептом, который мы обозначили словом *disappointed* (и которое соответствует также нем. *enttäuscht* и франц. *deçu*), состоит в том, что русское *разочарование* включает изменение фрагмента картины мира. При этом не обязательно, как это может показаться, речь идет о чем-то «высоком»: разочарование в любви, в людях, в жизни, в идеалах. Можно разочароваться в мобильных телефонах фирмы Сименс, но это все равно означает некоторое изменение в картине мира, т. е. что человек рассматривает данное обстоятельство как нечто существенное, важное (и это имплицировано употреблением слова *разочарование*). Естественно, возможна ирония, но это нисколько не противоречит сказанному. И наоборот, даже если идет речь об очень глубоком и сильном чувстве по важнейшему поводу (как в примере из «Саги о Форсайтах»), но оно ограничено реакцией на единичное событие и не включает никакой переоценки, то слово *разочарование* оказывается неуместно.

Итак, осталось понять, каков механизм семантической эволюции, приближающей сегодня специфическое русское *разочарование* к «банальному» европейскому. Семантическое калькирование с английского безусловно имело здесь место. Но мы можем показать, какие внутрисистемные факторы действующие в русском языке, способствовали такой семантической эволюции.

Мы выделили два класса отклоняющихся от строгой нормы употреблений; укажем, для каждого из них, факторы, обеспечивающие возможность такого отклонения.

Относительно употребления слова *разочарован* в контексте предикатных имен и с изъяснительным придаточным (*разочарован ее отказом/что она отказалась*) нужно сказать следующее. В русском языке имеется тенденция, состоящая в том, что предикаты, обозначающие эмоции, развивают оценочное или даже просто ментальное значение, из которого эмоция «выветривается»: таковы, например, глаголы *бояться* (⇒ ‘считать вероятным’), жалеть (⇒ ‘считать, что лучше было поступить иначе’), подозревать (⇒ ‘считать вероятным’). Кстати, здесь тоже происходит дополнительное упрощение семантики, в результате чего возникает квазисинонимия, ср. *я боюсь, что он вовремя не вернется* и *я подозреваю, что он вовремя не вернется*.

Особенно характерно здесь появление конструкции с придаточным с союзом *что* со значением факта. Это весьма закономерное явление, характеризующее эволюцию предикатов эмоционального состояния в сторону

ослабления эмоционального и усиления ментального компонента: такие глаголы как *бояться, надеяться, ждать, нравиться, жалеть, раскаиваться, волноваться, беспокоиться* и др. все в той или иной степени развивают это ментальное (оценочное, рациональное, в общем не собственно эмоциональное) значение, и оно реализуется именно в конструкции с *что*. Поэтому семантическое развитие предикатива *разочарован*, при котором он приобретает чисто оценочное значение и управление придаточным с *что*, является вполне закономерным.

Что касается второго класса употреблений (*это разочаровывает*), в чисто оценочном стативном значении *разочаровывает* встраивается в антонимическую пару к *радует* (они выражают, соответственно, отрицательную и положительную оценку; о таком употреблении глагола *радовать* (перфективный коррелят — *порадовать*) см. Зализняк, Шмелев 2003), ср.:

(42) *Проходя тестирование, вас могут* ***разочаровывать*** *или, наоборот,* ***радовать*** *полученные результаты.*

(43) *Петербургские баскетболисты* ***разочаровывают****, а хоккеисты* ***радуют*** *болельщиков.*

(44) *Мобильное телевиденье* ***радует*** *и* ***разочаровывает***

Тем самым *разочаровывает* в обсуждаемом употреблении не только не чуждо русскому языку, а просто необходимо ему, так как заполняет имеющуюся лакуну, т. е. его появление обусловлено именно системными факторами. В том же значении употребляются и некоторые другие глаголы: *удивляет, впечатляет, внушает оптимизм* и т. п.

Существенно, что то же значение, в соответствии с семантикой перфектной пары, возможно и в сов. виде — *разочаровал*; ср., соответственно, *порадовал*. Причем характерно именно такое абсолютивное употребление (без указания на субъект), подчеркивающее чисто оценочное значение, ср.:

(45) *«Демон» Рубинштейна* ***разочаровал*** *не по причине плохой постановки, а по причине отсутствия таковой: опера давалась в концертном исполнении. [Белла Езерская. Мариинский на сцене мет (2003) // «Вестник США», 2003.08.20]*

(46) *SonyEricsson K850i — Может, и хороший, но немного* **разочаровал**.

(47) ***Порадовал*** *отечественный концерн «Созвездие»…*

(48) *Но сегодня on-line словарь от google* ***порадовал!***

Итак, слово *разочарование*, исторически будучи словообразовательной калькой с франц. *désenchantement*, претерпело в русском языке определенную

семантическую эволюцию, в результате которой оно существенно разошлось с современным «общеевропейским» концептом, соответствующим значению слов англ. *disappointment*, франц. *déception*, нем. *Enttäuschung* и др., обозначающих чувство обманутого ожидания, вызванное оценкой некоторого внешнего по отношению к субъекту события. Между тем русское *разочарование* описывает событие внутреннего мира: состояние, вызванное фактом ***утраты ценности***, который видоизменяет ***индивидуальную картину мира***.

Семантический сдвиг (плохо идентифицируемый носителями языка, особенно принадлежащими к молодому поколению), демонстрируемый употреблениями типа *разочарован ее отказом/тем, что она отказалась/что она отказалась*, а также конструкцией *это разочаровывает*, очевидно, произошел в результате влияния английского языка, столь характерного для сегодняшней языковой ситуации. Однако скорее всего здесь действуют одновременно оба механизма, которые совпали по своему эффекту: семантическое калькирование, т. е. заимствование производного значения, пусть и очень недалеко отошедшего по пути семантической деривации, и независимая параллельная семантическая эволюция, определяемая сугубо внутрисистемными факторами (ср. Зализняк 2006: 415). Так, приобретение способности подчинять изъяснительное придаточное со значением факта — закономерное явление, характеризующее эволюцию предикатов эмоционального состояния: ср. глаголы *бояться, жалеть, ждать, нравиться, беспокоиться, волноваться, переживать* и др.). Поэтому семантическое развитие «в ментальную сторону» предикатива *разочарован*, при котором он приобретает управление придаточным с *что* и способность подчинять факт, является вполне закономерным и для своего объяснения не требует с необходимостью апелляции к иноязычному влиянию. Не обязательно калькированным оказывается при ближайшем рассмотрении также употребление *это разочаровывает*. Это выражение, конечно, является эквивалентом английского *it is disappointing*, однако внутри русского языка вполне активен механизм восстановления активной формы глагола эмоционального воздействия, ср. *это радует, удивляет, успокаивает* (ср. образованные по той же модели в языке молодежного сленга формы типа *меня прикалывает* или даже *меня улыбает*).

Литература

1. *Виноградов В. В.* История слов. Москва, 1994.
2. *Зализняк Анна А.* Семантика глагола бояться в русском языке // Известия АН СССР, сер. лит. и яз., т. 42, N°1, 1983. С. 59–66.
3. *Зализняк Анна А.* Многозначность в языке и способы ее представления. Москва: Языки славянских культур, 2006.
4. *Зализняк Анна А., Шмелев А. Д.* Homo jubilans // Rusistika. Slavistika. Lingvistika. Festschrift fur Werner Lehfeldt zum 60. Geburtstag. Herausgegeben von S. Kempgen, U. Schweier und T. Berger (Die Welt der Slaven, Sammelbande. Bd. 19). Munchen: Otto Sagner Verlag, 2003. S.310–319.

5. *Иорданская Л. Н.* Попытка лексикографического толкования группы русских слов со значением чувства. // Машинный перевод и прикладная лингвистика. Вып. 13. М., М: МГПИИЯ им. М. Тореза,1970.
6. *Мельчук И. А., Жолковский А. К.* Толково-комбинаторный словарь русского языка. Wiener Slavistischer Almanach. Sbd. 14, Wien, 1984.
7. *Черных П. Я.* Историко-этимологический словарь современного русского языка. В 2-х тт. М., 1994.
8. *Goddard C.* Semantic Analysis. N.Y.: Oxford Univ. Press, 1998.
9. *Iordanskaja L., Mel'čuk I.* Semantics of two emotion verbs in Russian: bojat'sja "to be afraid" and nadejat'sja "to hope" // Australian Journal of Linguistics, vol.10, № 2, 1990.
10. *Searle J. R.* Intentionality. An essay in the philosophy of mind. Cambridge et al, Cambridge Univ. Press, 1983.
11. *SOED* — Short Oxford English Dictionary. Fifth edition. Oxford: Oxford University Press, 2002.
12. *Wierzbicka A.* Negacja — jej miejsce w strukturze głębokiej. // A. Wierzbicka. Dociekania semantyczne. Wroclaw: Ossolineum, 1969.
13. *Wierzbicka A.* Semantic primitives // Linguistische Forschungen. B. 22. Frankfurt/M., 1972.
14. *Wierzbicka A.* The Semantics of Grammar. Amsterdam, 1988.
15. *Wierzbicka A.* Emotions across Languages and Cultures. Cambridge Univ. Press, Edition de la Maison des Sciences de l'homme. Paris, 1999.

СЛОВОСОЧЕТАНИЯ-ЗАПИСАННЫЕ-ЧЕРЕЗ-ДЕФИС КАК СПОСОБ КАТЕГОРИЗАЦИИ И НОВЫЙ ПРИЕМ ЯЗЫКОВОЙ ВЫРАЗИТЕЛЬНОСТИ В РУССКОМ ЯЗЫКЕ

Занегина Н. Н. (zanegina@gmail.com)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Статья посвящена одному из новых средств языковой выразительности и новому для русского языка способу словообразования, а главное — новому способу категоризации действительности, при котором письменная форма этого слова (или словесного образования) создается соединением с помощью дефисов простого/составного словосочетания или целого предложения (ср. «позвали женщину-для-уборки»). В результате анализа собранного в блогах корпуса примеров была создана классификация подобных образований с точки зрения причин их появления в тексте с учетом типа породившей их синтаксической структуры. Выяснилось, что данные образования обладают следующими функциями: 1) логическое выделение, подчеркивание определенного смысла, который обычно передается несколькими словами; 2) заполнение лексической лакуны; 3) выделение особого класса предметов; 4) цитирование прямой речи как типичной реакции на предмет/явление либо как ярко иллюстрирующей описываемое событие/явление; 5) подчеркивание связи зависимых слов именно с данным главным словом и некоторые другие. Рассматриваемые конструкции сопоставляются с такими, в которых 1) все слова образования пишутся слитно, 2) все слова соединены нижним подчеркиванием (_), 3) каждое слово пишется с прописной буквы. Их функции частично совпадают с функциями описываемых образований.

Ключевые слова: составные слова, дефисы, категоризация, языковая выразительность

1. Общие замечания

В русском языке дефис между словами используется для соединения дублирующихся слов (*давай-давай*) или слов, описывающих ряд действий/предметов (*им хочется погулять-посмотреть-попрыгать-побегать*). Однако в письменной интернет-коммуникации в последнее время весьма активно используются слова[1], образованные как соединение в единое целое простого/

[1] Термин «слово» применим к ним, поскольку они обладают смысловым единством и цельнооформленностью, а также выступают в предложении как одна синтаксическая единица.

составного словосочетания или целого предложения. Графически это соединение отображается с помощью дефисов:

(1) *позвали женщину-для-уборки*.

Большинство подобных конструкций встречается в блогах и им подобных типах текстов, однако аналогичные конструкции были замечены в текстах СМИ[2], художественных произведений[3] и даже научных текстах в качестве вводимого автором термина[4].

Соединение слов в одно большое новое слово возможно тремя способами: с помощью дефиса, с помощью знака «нижнее подчеркивание» (_) и за счет написания этой конструкции слитно, без пробелов. В анализируемом здесь корпусе примеров наиболее частотными оказались дефисные конструкции. Действительно, конструкции без пробелов трудны для восприятия (ср.: (2) *а в договоре это прописано, нуактожечитаетдоговор*), а нижнее подчеркивание не воспринимается как знак препинания, который может находиться внутри слова (ср.: (3) *Мне бы понять, правда ли одинокий пожилой человек может_так_подумать*…). Также типологически близкими являются конструкции, в которых каждое слово пишется с прописной буквы ((4) *ко мне пришла подружка со своим мужем — Очень Творческим Человеком*).

Необходимо учесть, что в английском языке существуют так называемые составные слова, которые обычно образуются соединением трех основ, графически оформленных дефисами (*easy-to-use*); чаще всего являются прилагательными. Такие слова могли оказать влияние на возникновение дефисных конструкций в русском языке, однако этот вопрос требует дополнительного рассмотрения.

В английском языкознании[5] такие конструкции называют свернутыми (термин предложен в работе [Петрова 2004]). На данный момент установлено, что эти конструкции являются закономерным развитием языка аналитического типа, обладают несколькими десятками моделей построения, но при этом часто образуются окказионально, для придания тексту большей экспрессивности, а главное — служат категоризации действительности [Там же][6].

---

[2] «Когда невеста уходит, Таня с видом ну-я-же-говорила поворачивается ко мне» («Русский репортер», http://rusrep.ru/article/2011/12/13/flowers).

[3] «Славы Моцарта-от-науки юный сибиряк не стяжал, да он, правду сказать, к ней и не стремился» (Б. Акунин. Квест).

[4] Ср. особый тип писем, который К. А. Гилярова выделяла в переписке между студентами и преподавателями и назвала «письма-по-существу» [Гилярова 2011].

[5] Ср. посвященные им работы [Букина 2002], [Петрова 2004].

[6] Историю возникновения таких конструкций в русском языке пока не удалось проследить.

В русском языке категоризация и экспрессивность также являются функциями дефисных образований, но существуют и некоторые другие.

2. Типы дефисных образований

Для данного исследования методом сплошной выборки был собран корпус минимальных контекстов, в которых встречаются дефисные образования. Источник контекстов — записи в «Живом журнале», сделанные в 2011–2012 году более чем 30 авторами (все — женщины). На данный момент обработано более 120 контекстов. Такое гендерное единообразие респондентов отчасти случайно (при сплошном просмотре текстов блогов искомые дефисные образования нашлись пока только в блогах женщин), отчасти неслучайно, поскольку если бы эти образования использовались мужчинами и женщинами одинаково часто, то среди 30 авторов должны были бы встретиться авторы-мужчины.

В исследуемом корпусе контекстов удалось выделить как минимум пять типов дефисных образований, которые условно можно обозначить следующим образом: «Тот самый этот-предмет», «Не-знаю-как-назвать-предмет», «Не-хочу-называть-предмет», «Такого-слова-почти-нет», «Предмет я-обычно-о-нем-говорю-так».

2.1. «Тот самый этот-предмет»

Одна из основных причин появления неологизмов — необходимость поименовать новое явление действительности, с которым столкнулось общество. Окказионализмы (авторские неологизмы) возникают тогда, когда с чем-то новым для себя сталкивается отдельный человек. При этом важное отличие неологизмов, передаваемых дефисными написаниями, от неологизмов, передаваемых на письме любым другим способом, заключается в том, что первые именуют понятия, которые уже имели к этому моменту описательное название, однако в языковой картине мира не были выделены в отдельную категорию, требующую специального отдельного наименования или имеющую право на такое наименование. Ср.:

(5) *Согласовала с мужчиной идею «завести» женщину-для-уборки.*

«Женщина-для-уборки» отличается от «женщины для уборки» тем, что в первом случае автор как бы активизирует в сознании читателя весь комплекс признаков концепта «женщина-для-уборки»: периодически приходящая помощница, которая за определенную плату убирает квартиру. Синонимичные этому образованию «уборщица», «домработница», «приходящая домработница» кажутся автору недостаточно отражающими реальность, может быть, слишком широким (это не домработница, это человек, выполняющий одну

конкретную часть ее типичных функций) либо обладающим ненужной отрицательной коннотацией.

Ср. аналогичные конструкции:

(6) *и вот про еду расскажите — её всё-таки на весь полет туда-обратно запасают (например, в Гонконг), или в зависимости от не-знаю-чего?*

(7) *Мама выхудела из магазинов-для-толстых, но не вхудела в размерную цепочку большинства обычных магазинов.*

Функцию дефисов в этих примерах можно сопоставить с функцией определенного артикля и кратко описать этот тип дефисных образований как называющие ***«Тот самый этот-предмет»***. Также часто в этом типе сочетаний

Частные случаи этого типа использования дефисных образований:

— подчеркивание наличия логического ударения именно на этом сочетании слов в предложении:

(8) *Я в жж к ней иногда хожу просто набраться спокойствия и позитива-в-хорошем-смысле.*

— устранение двусмысленности в синтаксических связях слов в предложении:

(9) *А это гендер) Мужчине нужен Результат-любой-ценой)))* (автору важно, чтобы при прочтении была выделена не синтагма «нужен любой ценой», а синтагма «результат любой ценой»; впрочем, задача выделения нового типа явления — именно так достигнутого результата — здесь также сохраняется);

— устранение двусмысленности вместе с акцентуацией атрибутивного признака присоединенных дефисом единиц:

(10) *Поздравляем! Ваш сын зачислен в школу-по-прописке!* (самая важная характеристика этой школы — не номер, не ее специализация, а то, что она приписана к такому-то адресу прописки).

2.1.1. Отступление о механизме создания именования

В процессе поиска сочетания слов, которые можно «склеить» в дефисное образование, автор может пользоваться как стандартными синтаксическими конструкциями, не нарушающими грамматические правила и могущими употребляться в такой грамматической форме и без дефисов, так и такие, которые претерпели определенные грамматические изменения.

Стандартные в свою очередь могут быть разделены на

— восходящие к словосочетанию (см. примеры 7, 10) и

— восходящие к целому предложению, в том числе сложному:

(11) *У меня все мегапросто: много много работы, мало мало сна, между ними влезает еда, обязательная помощь близким, обязательные встречи на-которые-нельзя-не*[7] *придти, алкоголь и книжки в метро;*

(12) *Дамы, а расскажите историй про то, как вы или кто-то «справился» с человеком-который-знает-как-надо, или договорился с ним о границах, или еще как-нибудь уменьшил разрушительное действие.*

Грамматически дефектные конструкции получаются за счет:
— пропуска одного слова:

(13) *череда неприятных вопросов-в-лоб* (вопросов, заданных в лоб);

— пропуска нескольких слов:

(14) *у меня вот не-муж-но-тем-не-менее раньше поедал все подряд целыми днями, потом решил, что поправился, и сел на диету*[8] (не муж, но тем не менее значимый для меня человек);

— трансформации устойчивой конструкции:

(15) *Когда генные инженеры выведут породу собак-до-старости-щенков … — все равно жизненный опыт и годы будут видны в выражении глаз* (маленькая собака до старости щенок).

2.2. «Не-знаю-как-назвать-предмет»

Более простой случай — отсутствие не только точного наименования для какого-то типа предметов, но и синонимов, которые могли бы назвать этот предмет хотя бы приблизительно. Поиск нового слова разрешается соединением старых слов при помощи дефиса с одновременным признанием невозможности подобрать одно точное слово. Заодно автор решает проблему

---

[7] Неприсоединение к дефисному образованию последней потенциальной его части (*встречи на-которые-нельзя-не<->придти*) может быть как ошибкой автора, так и сознательным приемом, поскольку такое одиночное использование отрицательной частицы часто встречается как пример эмоционально маркированной экономии речевых средств, когда более или менее очевидное слово, которое должно было бы следовать за частицей «не», опускается и тем самым акцентируется, одновременно придавая тексту интонацию доверительности, интимного умолчания, ср.: *чего я только уже не, а все без толку.*

[8] Поиск слова для обозначения партнеров в новых отношениях семейного типа довольно часто приводит к образованию новых дефисных конструкций: (11) *Летом мы с моим без-пяти-минут-мужем придумали развлечение…;* (12) *У меня сегодня ровно год, как меня зафрендил мужчина-с-которым-я-просыпаюсь-уже-десять-месяцев.*

объяснения значения придуманного им слова: объяснение содержится в самом именовании.

(16) *можно было брать напрокат велосипеды и не-знаю-как-обозвать-эти-штуки-с-четырьмя-колесами*

(17) *купила юбку не-знаю-как-называется-волнами-от-талии-изображает-офисную*

2.3. «Не-хочу-называть-предмет»

Близкий вариант — нежелание автора вспоминать, уточнять или просто использовать какое-либо слово, которое читатель скорее всего вспомнит сам. Обычно этим словом оказывается имя собственное.

(18) *Дамы, а что это вы так претесь все по этому Бенедикту Как-его-там-с-непроизносимой-фамилией, который Шерлока играет?* (трудно вспомнить/произнести эту фамилию);

(19) *Жена летчика Ярошенко попросила помощи у сами-знаете-кого*[9] (не хочется лишний раз произносить эту фамилию);

(20) *В ответ на восторги двух-известных-блоггеров после поездки в город будущего Шанхай придумал хорошую шутку* (эвфемистичное сочетание может называть более одного человека);

(21) *Меня попросили прислать «красочную фотографию» для ежегодной университетской брошюры в связи с защитой — «просто портрет, и чем неформальнее, тем лучше». Как догадываются молодые мамашки, у меня не оказалось фотографий без сами-знаете-кого в кадре* (подразумевается ребенок).

2.4. «Такого-слова-почти-нет»

Данный тип образований объединяет в себе те единицы, которые восполняют словообразовательные лакуны и являются гибридом двух нормативных слов или словосочетаний:

(22) *по улицам ходят миллионы двадцати-с-небольшим-летних* (двадцатилетние + двадцать с небольшим);

[9] Из твитера пользователя lentaruofficial.

(23) *по результатам двух-с-половиной-часовой беседы написала статью* (двухчасовой + два с половиной).

2.5. «Предмет я-обычно-о-нем-говорю-так»

Особый тип представляют дефисные образования, в которых цитируется чья-либо прямая речь. Эти слова самого автора либо слова других людей могут быть

— образцом типичной реакции на предмет/явление:

(24) *вообще все мои мелкие радости — из детства, и люблю я их, видимо, за то, что они дают мне ощущение детства и ощущение «мы-понимаем-друг-друга-без-слов»;*

(25) *Зачем? Муж до сих пор не понимает, не могу сказать, чтобы он сильно стремился понять. Скорее относит это к у-всех-же-должны-быть-странности. Но моральную поддержку оказывает;*

— яркой иллюстрацией описываемого события/явления:

(26) *Уже 2 или 3 раза он меня кидал, по телефону сообщая, что вот-уже-прямо-сейчас-едет в офис или завтра-прямо-с-самого-утра будет там с моими деньгами.*

Такое цитирование, помимо создания яркого образа, бытовой картинки, одновременно как бы снимает с автора ответственность за сказанное, за выбранные слова для описания явления, поскольку ответственность за эти слова перекладывается на саму действительность[10].

Частными случаями этого цитатного типа дефисных образований являются:

— цитирование многословного прецедентного текста:

(27) *Да, дочка уже после рождения выдавала классическую реакцию узнавания на «полковнику-никто-не-пишет», которую я слушала в промежутках между операми и балетами*;

[10] Ср. рассмотренный в работе [Иомдин, Бердичевский 2006] особый случай использования местоимения *этот* при имени собственном, когда говорящий с помощью такой лексемы показывает, что лишь повторяет номинации, использованные собеседником или другими людьми; см. также упоминаемой в этой работе явление косвенной эвиденциальности (пересказывательности), выражающейся во многих языках грамматически и описанной Н. А. Козинцевой. Благодарю анонимного рецензента за сообщение об этой работе, а также за ряд прочих важных замечаний, которые, к сожалению, не все могли быть учтены в данный момент.

— использование прямой речи для именования нового типа вещей:

(28) *для иллюстрации ух-ты-эффекта использую…*

(29) *Не знаю, что мною двигало <…> когда я покупала эту посмотрите-вот-моя-грудь футболку.*

3. Всесловаслитно

Написание всех слов образования без пробелов между ними к цели категоризации действительности добавляет оттенок значения «замечание в скобках» или «замечание в сторону», предполагающее ускоренную интонацию при произнесении всей конструкции.

(30) *а в договоре это прописано, нуактожечитаетдоговор* (в одно слово соединена часто повторяющаяся мысль)

У таких образований выше, чем у дефисных, степень типичности, формульности соединенного в одно слово образования.

Для усиления этой формульности после такого образования часто ставится знак копирайта или торговой марки (©, (c), ™, ТМ):

(31) *я таки уехала, и даже с самойбольшойлюбовьювмоейжизни ТМ потом ни раз встретилась, но только ничего у нас не было.*

Эта формульность может стать предметом языковой рефлексии:

(32) *— вот это вот «уменяжеребенок» — это вообще уникальная формула для многих, похоже… <…> а чо, у меняжетожеребенок, мне его кормить надо*

4. «Выделяем очень_важное_понятие»

Использование знака «нижнее подчеркивание» для соединения составных образований чаще является маркером логического ударения, которое иначе можно было бы передать на письме подчеркиванием всего словосочетания, но это не всегда возможно в интернет-коммуникации по техническим причинам[11].

[11] В некоторых редакторах нижнее подчеркивание как технический знак ставится в начале и конце слова или ряда слов, которые необходимо выделить каким-либо шрифтом.

(33) *Мне бы понять, правда ли одинокий пожилой человек может_так_подумать...*

(34) *банкетный зал — это что-то слишком пафосное, как для свадьбы. я надеялась на просто_место.*

5. «Так выделяются Самые Важные Слова»

Несколько иначе устроено использование всех прописных букв в простом или составном словосочетании. Здесь автор не настаивает на том, что все слова вместе представляют цельное понятие, а обращает внимание на значимость каждого слова из написанных с прописной. Наиболее частотная функция такого написания — назвать такой признак или явление, в котором в данном конкретном случае максимально выраженны типичны характеристики, свойственными этому признаку или явлению:

(35) *И вот приходит мне письмо с Работы Моей Мечты* (это настоящая работа мечты, та самая, единственная);

(36) *У меня была книга, которой ни у кого не было. Важная и Нужная Книга* (очень важная и очень нужная, практически единственная);

(37) *ко мне пришла подружка со своим мужем — Очень Творческим Человеком* (муж обладает всеми основными признаками очень творческого человека).

Знак копирайта или торговой марки также часто сопутствует подобным сочетаниям:

(38) *Один мальчик так куртуазно выражается, что решись я цитировать его в жж, то мне бы тут же посоветовали лечить его от аутизма и шизофрении, потому что Нормальные Дети ТАК Не Говорят! (тм).*

Итак, условно можно выделить следующие пять типов дефисных образований: «Тот самый этот-предмет», «Не-знаю-как-назвать-предмет», «Не-хочу-называть-предмет», «Такого-слова-почти-нет», «Предмет я-обычно-о-нем-говорю-так». Естественно, что между описанными конструкциями не всегда существуют жесткая граница. Чтобы установить, насколько универсальны правила соединения таких дефисных образований, было бы интересно провести следующий эксперимент: попросить информантов найти слова в текстах разного типа (публицистическом, поэтическом, повествовательном), которые, по их мнению, могли бы быть соединены в дефисные образования, и сопоставить процент совпадающих выборов.

Эти дефисные образования могут выполнять в тексте следующие функции: 1) логическое выделение, подчеркивание определенного смысла, который

обычно передается несколькими словами; 2) заполнение лексической лакуны; 3) выделение особого класса предметов; 4) цитирование прямой речи как типичной реакции на предмет/явление либо как ярко иллюстрирующей описываемое событие/явление; 5) подчеркивание связи зависимых слов именно с данным главным словом и некоторые другие.

Литература

1. *Букина В. А.* (2002) Английское сложное прилагательное с дефисным написанием и его роль в тексте. Автореф. дисс… канд. филол. наук. М.
2. *Петрова Т. А.* (2004) Свёрнутые конструкции как тип сложных номинативных единиц в современном английском языке. Автореф. дисс… канд. филол. наук. Иркутск.
3. *Гилярова К. А.* (2011) Курсач в аттаче: особенности электронной коммуникации между преподавателем и студентом // http://www.dialog-21.ru/digests/dialog2011/materials/ru/pdf/24.pdf
4. *Иомдин Б. Л., Бердичевский А. С.* А кто этот этот? Имена собственные и неопределенная определенность // http://www.dialog-21.ru/digests/dialog2006/materials/html/IomdinB.htm.

TOWARDS A SYSTEM OF SYNTACTIC DEPENDENCIES OF GERMAN

Zangenfeind R. (r.zangenfeind@lmu.de)

University of Munich, Munich, Germany

A set of syntactic relations in a dependency grammar for German is proposed which can be used for machine translation and other applications. The basis for this set is the formal model of Russian syntax used in the linguistic processor ÈTAP, which is built on the basis of Meaning <=> Text theory. Using a terminology as close to ÈTAP as possible facilitates a potential computational implementation. At this stage of development the system of German dependency syntax comprises 58 syntactic relations, which were used to analyze manually several hundred German sentences. 18 of them have definitions that are similar to those of their counterparts in Russian syntax. Seven relations have the same definitions as their Russian counterparts except for the concrete German lexemes that are part of the definitions. For another 30 relations identical definitions as for the corresponding Russian relations can be used. Three German relations do not have a Russian counterpart.

Key words: surface syntactic relation, dependency, machine translation, Meaning <=> Text theory, German, Russian

Introduction

In this paper a first sketch of a set of surface syntactic relations in a dependency grammar for German is proposed, that can be used for machine translation. The basis for this set was the formal model of Russian syntax that was developed by Ju. D. Apresjan, I. M. Boguslavskij, L. L. Iomdin and others for the linguistic processor ÈTAP-3 (cf. Apresjan et al. 2010), which is built on the basis of Meaning <=> Text theory (cf. Mel'chuk 1974). This model of Russian syntax is an excellent starting point because it is field-tested in the practice of NLP for a long time and still is continued to be refined (cf. Boguslavsky et al. 2011).

An essential part of ÈTAP is the rule-based machine translation from Russian to English and vice versa (cf. Apresjan et al. 1989 and Apresjan et al. 2003). The machine translation from Russian to German is an additional part of ÈTAP which has a prototypical character until now. My aim is to develop further this part and as a first step I want to propose a system of syntactic relations for German. This should be as similar as possible to the system of Russian relations in ÈTAP to have a basis for a computational implementation of Russian-German translation. Apresjan et al. (2010:24–43) describe seventeen actantial relations, 32 attributive, five coordinative and eight auxiliary relations for Russian. For the names and definitions of German relations I followed in a first step, whenever possible, the Russian and English

relations that are used in ÈTAP (cf. Apresjan et al. 2010:24–43 and Apresjan et al. 1989:71–121), and then adapted them to the requirements of German syntax.[1] German linguists, like e. g. Heringer (1996), Engel (2009) or Kunze (1975), often do not use any special labels for syntactic relations,[2] whereas Specht (2003) uses labels that are partially different from ÈTAP's syntactic relations. I, nevertheless, will use a terminology as close to ÈTAP as possible in order to facilitate a computational implementation. In a large part the definitions of Russian and German relations are more or less the same or at least similar. But there are some relations in German that do not exist in Russian, whereas some Russian relations have no counterparts in German,[3] and there are differences, of course, when German lexemes (or Russian lexemes respectively) are part of the definition. Besides the possible practical use, this system of syntactic relations might make a little contribution to the theory of translation.

The syntactic relations presented below were used to analyze manually several hundred German sentences, most of which are translations of Russian sentences that represent a corpus including all the different syntactic relations of Russian.[4] This, of course, is just a start; so, it is possible that one or another additional relation might be needed when a larger corpus with more different German constructions will be analyzed in future work.

The paper is structured as follows: German relations with definitions different from their Russian counterparts are presented in the first part; they are of particular importance for translation: actantial relations in section 1, attributive relations in section 2 and auxiliary relations in section 3 (there are no coordinative relations in this part). Relations with definitions that are more or less the same as those of their Russian counterparts are found in section 4. The paper ends with conclusions in section 5.

1. Actantial Syntactic Relations

1.1. The **prädikative syntaktische Relation** 'predicative syntactic relation'[5] connects a personal verb as syntactic governor with the grammatical subject of the

---

[1] I want to use German names for German relations in order to take into account their language-specific character.

[2] Kunze (1975:19) uses so called *Markierungen* 'marks' to characterize dependent nodes, like e. g. „temporaler Akkusativ" 'temporal accusative' or „Akkusativobjekt" 'accusative object'. Heringer (1996:36) uses lexical categories to characterize nodes in dependency trees (stemmas), like e. g. „V" (for verbs) or „PTL" (for particles). Engel (2009:305–309) characterizes nodes with labels like „E_{sub}" (for subject) or „Amod" for (Modifikativangabe 'modifying phrase'), which comes at least near labeling relations.

[3] Seven out of the 62 Russian relations that are described by Apresjan et al. (2010:24–31) do not have a counterparts in German, cf. Zangenfeind (2011:313f.) for considerations in the field of actantial relations.

[4] The Russian sentences were taken from Apresjan et al. (2010), Apresjan et al. (1992) and from http://www.ruscorpora.ru; original German sentences were taken from Süddeutsche Zeitung.

[5] I will provide English translations for the labels of German relations only where it might be needed.

phrase as dependent. The subject can have one of the following forms: a noun in the nominative case, a preposition (representing a prepositional group), an adverb, an adjective, the particle *zu* 'to' (with an infinitive verb), a conjunction (introducing a relative clause) or a personal verb (representing a relative clause); cf. the following example, where X marks the syntactic governor and Y the dependent:[6]

(1) *Er [Y] liest [X] ein Buch*. 'He [Y] reads [X] a book'

The definition of this German relation differs from the corresponding Russian relation in some details: the syntactic governor of this relation in Russian can also be a noun, an adjective, the short form of a participle or an adverb in constructions without copula verb, which is not possible in German, unless in elliptical constructions. The dependent, i. e. the subject, in Russian can also be a noun in the genitive or partitive case, the infinitive verb (in those cases where in German the particle *zu* 'to' is the dependent), which, again, is not possible in German; cf. the definition of the Russian predicative relation by Apresjan et al. (2010:25). A further difference between the German *prädikative Relation* and the corresponding Russian relation consists in different requirements of agreement between verb and subject.

1.2. The **agentive SyntRel** connects the predicate (the participle part of a verb in passive voice or an attributive passive participle) as syntactic governor with the preposition *von, durch* 'by, from', *seitens* or *von Seiten* 'on the part of', which introduces the semantic subject of an action denoted by the predicate (this subject corresponds to the first semantic actant of the predicate in active voice):

(2) *Das Ergebnis wurde von [Y] Experten bestätigt [X]*. 'The result was confirmed [X] by [Y] experts'

The Russian agentive relation is used both for verbs in passive voice and for nouns as syntactic heads, if the dependent is a noun in the instrumental case, cf. the Russian definition by Apresjan et al. (2010:26); I propose to use the German *quasi-agentive Relation* (cf. section 1.3) in all such cases where a noun is the syntactic head, cf. Zangenfeind (2011:311). The dependent of the Russian agentive relation is always a noun in the instrumental case.

1.3. The **quasi-agentive SyntRel** connects a predicative noun with its first semantic actant, which is a noun in genitive case or a prepositional group that is introduced by the preposition *von, durch* 'by, from', *seitens* or *von Seiten* 'on the part of':

(3) *Das ist ein Bericht [X] von [Y] unserem Kollegen*. 'This is a report [X] by [Y] our colleague'

[6] Due to lack of space, I will give only one example for each relation.

1.4. The **kopulative SyntRel** connects the copula (*sein, bleiben, werden* 'to be, to stay, to become') with the nominal part of the complex predicate. The dependent can have one of the following forms: a noun in the nominative case, an adjective or an adverb, a prepositional group or a group that is connected via conjunction:

(4) *Er ist [X] krank [Y].* 'He is [X] sick [Y]'

In addition to the Russian equivalents of the three German copula verbs, in ÈTAP four more Russian verbs are marked with the syntactic feature 'СВЯЗ' as copula verbs, namely *бывать* 'to be, to visit, to happen, to take place', *делаться1* 'to become', *казаться* 'to seem', *оказываться* 'to turn out to be'; they all can be the governor of the corresponding Russian copulative relation, which is not possible in German. The dependent of the Russian relation can also be a word in the instrumental case or in the genitive case.

Another point is the following: in the phrase *Er war [X] in [Y] Bulgarien* 'He was [X] in [Y] Bulgaria' (cf. a similar Russian example by Apresjan et al. (2010:27)) *sein* 'to be' is used in the meaning of 'to be located'; so, maybe it should be reconsidered, whether here we don't have a *kopulative Relation*, but a *1. kompletive*, which connects a full verb with its the first complement.

1.5. The **kompletiv-appositive SyntRel** connects a noun that denotes e. g. a parameter with its actant if it is a noun in the nominative case that denotes a quantity:

(5) *Diamant hat die Härte [X] 10 [Y].* 'Diamond has the hardness [X] 10 [Y]'

The definition of the corresponding Russian relation is very similar; the dependent, however, can also be an adverb with quantitative meaning etc.

1.6. The **komparative SyntRel** connects a word with comparative meaning or a verb, a noun or an adverb with a comparative conjunction (e. g. *wie* 'like' or *als* 'as') that introduces the second part of the comparison:

(6) *Es ist kürzer [X] als [Y] ein Meter.* 'It is shorter [X] than [Y] one meter'

The definition of the corresponding Russian relation, again, is very similar; but here, in Russian the dependent additionally can be a noun in the genitive case. The concrete conjunctions that are used as a dependent element in the corresponding Russian relation are, of course, different. This also holds for all subsequent relations when concrete German lexemes are part of a definition.

Another point is the following: it might be useful to split this relation into two relations (one for syntactic governors that have a comparative meaning and another for verbs, nouns and other words that don't have this comparative meaning), because when the syntactic governor is a word without comparative meaning the definition does not match actantial relations but attributive relations, cf. example 7:

(7) *Ich vertraue [X] ihm wie [Y] einem Freund.* 'I trust [X] him like [Y] a friend'

1.7. The **elektive SyntRel** connects a word that denotes a choice out of a quantity (i. e. a numeral or an ordinal number, an adjective in comparative or superlative degree) with one of the prepositions *von, aus, unter* 'of, out of, between' or a noun in the genitive case:

(8) *Der bessere [X] der beiden Vorträge [Y]. 'lit*. The better-one [X] of-the two talks [Y]'

The syntactic governor of the corresponding Russian relation can also be one of the words that are used to build analytical superlative forms of adjectives.

2. Attributive Syntactic Relations

2.1. The **(eigentliche) modifikative SyntRel** 'modifying relation proper' connects a noun, an adjective, an ordinal number or a participle that is used as an attribute X with an adjective or the like Y that agrees with X in gender, number, case and definiteness:

(9) *Ein schöner [Y] Baum [X].* 'A beautiful [Y] tree [X]'

Here, comparing the German and the corresponding Russian relations, we have only different requirements of agreement between syntactic governor and dependent: in Russian, definiteness is no criterion, but on the other hand, animacy is an additional criterion.

2.2. The **determinative SyntRel** connects a noun with its determiner. Heringer (1996:59) lists the following groups of determiners for German: definite: *der, die, das* 'the'; indefinite: *ein, eine*, ... 'a'; demonstrative: *diese, jene*, ... 'these, those'; quantitative: *alle, jeder*, ... 'any, every'; negative: *kein, keine,* ... 'no'; possessive: *mein, ihr,* ... 'my, her'; interrogative: *welche* 'which'. An example with definite determiner:

(10) *Der [Y] Baum [X].* 'The [Y] tree [X]'

This is a relation that does not exist in Russian. A corresponding relation, nevertheless, is used for English, cf. Apresjan et al. (1989:99f.).

In the combination determiner + adjective + noun (cf. example 9) Engel (2009:52f.) treats the determiner as syntactic head of the adjective, because the declension of the adjective is dependent on the determiner (which represents morphological dependency), whereas Kunze (1975:65f.) uses different criteria to show that both, adjective and determiner are direct dependents of the noun; Heringer (1996:247) also treats adjective and determiner as direct dependents of the noun. I will follow Kunze's and Heringer's solution because this seems to better represent the syntactic circumstances.

2.3. The **(eigentliche) attributive SyntRel** 'attributive relation proper' connects a noun or an adjective with its non-congruent attribute (e. g. a noun in the genitive case, a prepositional group or an adverb):

(11) *Das Haus [X] vorne [Y] ist neu. 'lit.* The house [X] in-front [Y] is new'

The syntactic dependent of the corresponding Russian relation can also be a noun in the instrumental case or an adjective in comparative degree.

2.4. The **(eigentliche) appositive SyntRel** 'appositive relation proper' connects a noun as syntactic governor with an apposition:

(12) *Die Häuser der Stadt [X] Berlin [Y]. 'lit.* The houses of the city [X] Berlin [Y]'

In the corresponding Russian relation agreement between syntactic governor and dependent usually is required, which is not the case in German. Another difference between the Russian and the German appositive relations is the following: in Russian personal names the relation is directed from first to second name; in German there is reason to assume it is the other way round, cf. the genitive case of a personal name, which is marked only at the second name:[7]

(13) *Franz [Y] Kafkas [X] Werk.* 'Franz [Y] Kafka's [X] oeuvre'

2.5. The **quantitativ-koprädikative SyntRel** 'quantitative-copredicative relation' connects a verb with a quantitative group (e. g. a numeral) representing a co-predicative item. The noun that is described by the co-predicative item and separated from it is not necessarily in the genitive case as it is in the corresponding Russian relation but it is usually in the case that is required by the verb:

(14) *Bücher lieferte [X] man eine ganze Kiste [Y]. 'lit*. Books they delivered [X] a whole box [Y]'

2.6. The **(eigentliche) adverbiale SyntRel** 'adverbial relation proper' connects a verb with an adverbial phrase (i. e. an adverb, a noun in the accusative or genitive case, a prepositional group, an adverbial participle, a verb in the infinitive, a subordinate clause introduced by a conjunction):

(15) *Sie liest [X] schnell [Y].* 'She reads [X] fast [Y]'

The syntactic dependent of the corresponding Russian relation can also be e. g. a noun in the instrumental case or an adjective in the genitive case.

2.7. The **durative SyntRel** connects a verb that does not have any valency of duration with an adverbial phrase in the form of a nominal group in the accusative case that has the meaning of a duration:

[7] cf. also Helbig, Buscha (2001:511) and Heringer (1996:103).

(16) *Er spielt [X] den ganzen Tag [Y] Gitarre.* 'He plays [X] the guitar all day [Y]'

The syntactic dependent of the corresponding Russian relation can also be a prepositional group.

2.8. The **Distanz-SyntRel** 'distance relation' connects a verb with an adverbial phrase of spatial distance that is expressed by a noun in the accusative case:

(17) *Er ging [X] einen Kilometer [Y].* 'He went [X] one kilometer [Y]'

Here, again, the syntactic dependent of the corresponding Russian relation can also be a prepositional group.

2.9. The **adverbial-tautologische SyntRel** 'adverbial-tautological relation' connects a verb with a noun in the accusative case that duplicates part of the meaning of the verb:

(18) *Sie lebte [X] ein kompromissloses Leben [Y].* 'She lived [X] an all-out life [Y]'

The syntactic dependent of the corresponding Russian relation is a noun in the instrumental case.

2.10. The **Subjekt-koprädikative SyntRel** 'subject-copredicative relation' connects a verb with a co-predicative item that characterizes the subject of the verb and that is expressed by a prepositional group or an adverb:

(19) *Meine Schwester kam [X] mit [Y] gebrochenem Bein ins Krankenhaus.* 'My sister came [X] to the hospital with [Y] a broken leg'

The syntactic dependent of the corresponding Russian relation cannot be an adverb but a noun or an adjective in the nominative or instrumental case, which is in congruency in gender and number with the subject of the verb; (it also can be a prepositional group, as in German).

2.11. The **Objekt-koprädikative SyntRel** 'object-copredicative relation' connects a verb with a co-predicative item that characterizes the object of the verb and that is expressed by a prepositional group or an adverb:

(20) *Man lieferte [X] ihn sterbend [Y] ins Krankenhaus ein.* 'They took [X] him to to the hospital, dying [Y]'

The syntactic dependent of the corresponding Russian relation is a noun or an adjective in the accusative or instrumental case, which is in congruency in gender and number with the object of the verb (or a prepositional group, as in German).

3. Auxiliary SyntRel

3.1. The **analytische SyntRel** 'analytical relation' connects two elements of analytical verb forms (but not of passive voice); the inflected word form is the syntactic governor:

(21) *Wir werden [X] gehen [Y].* 'We will [X] go [Y]'

The syntactic dependent of the corresponding Russian relation can be a particle (Russ. *бы*) to build e. g. the subjunctive, which is not the case in German.

3.2. The **präfigurative SyntRel** 'prefigurative relation' connects a compound verb with its separable prefix:

(22) *Er montierte [X] die Antenne ab [Y]* [infinitive of the verb: *abmontieren*]. *'lit.* He mounted [X] the antenna dis [Y] [= He dismounted the antenna]'

This is a relation that does not exist in Russian. A similar, but in detail different relation is used for English phrasal verbs (phras-junct), cf. Apresjan et al. (1989:115f.).

3.3. The **reflexiv-analytische SyntRel** 'reflexive-analytical relation' connects a verb with its related reflexive particle *sich* 'oneself' and its word forms:

(23) *Sie verstecken [X] sich [Y]. 'lit.* They hide [X] themselves [Y]'

This is a relation that, too, does not exist in Russian. It is similar to the Russian relation *вспом* 'auxiliary' (cf. Apresjan et al. (2010:42)), but differs from it in that the German verb and its related reflexive particle *sich* 'oneself' together form a dictionary entry of its own.

4. German relations with the same definitions as their Russian counterparts

Some German relations have the same definitions as their Russian counterparts except for the concrete German lexemes that are part of the definitions instead of Russian lexemes in the corresponding Russian relations. In the field of attributive relations these are:[8]

- **verschoben-attributive SyntRel** 'displaced-attributive relation': the German preposition *bei* 'at' is the syntactic dependent of this relation to connect an external possessor, instead of the equivalent Russian preposition.

[8] Due to limited space I don't give the whole definitions here; cf. Apresjan et al. (2010:33, 35, 39, 41, 43). For the same reason no examples are shown.

- **quantitativ-restriktive SyntRel**: the dependent of this relation, that describes an intensity, can be connected via the German preposition *um* 'at'.
- **distributive SyntRel**: the dependent of this relation, that describes a unit of measure, can be connected via one of the German prepositions *in, pro, auf, je* 'in, per' or the like.
- **erklärende SyntRel** 'explaining relation': the subordinate clause (whose predicate is the dependent of this relation), that explains a fact of the main clause, is introduced with one of the German relative connectives *weshalb, was, womit, wodurch* 'which is why, what, with what' or the like.
- **erläuternde SyntRel** 'illustrative relation': the syntactic governor of this relation, whose dependent is a word referring to elements of a generic term, can be e. g. one of the German words *alles, überall* 'all, everywhere' etc.

in the field of coordinative relations it is the

- **sequentielle SyntRel** 'sequential relation': the syntactic dependent of this relation (a word with the same character as its governor) can be connected via one of the German prepositions *auf, mal, zu, gegen* 'by, times, versus'.

in the field of auxiliary relations it is the

- **expletive SyntRel**: As a syntactic governor there is a German word like *es, dies, jener, davon, das* 'it, this, that, of this, what' etc., that stands for another phrase in the clause, that is the dependent of this relation.

Moreover there are thirty German relations for which the same definitions as for their Russian counterparts can be used as long as no language-specific terms will be part of them. This might undergo a change when more details about these relations can be stated as a result of future work. Due to lack of space they are just listed here without definitions[9] and examples; in the field of actantial relations there are: **Dativ-Subjekt-SyntRel, uneigentliche agentive, erste kompletive** etc., **erste uneigentliche kompletive** etc., **nichtaktantisch-kompletive, präpositionale, subordinierend-konjunktionale, komparativ-konjunktionale SyntRel**; in the field of attributive relations: **deskriptiv-modifikative, relative, kompositive, isoliert-appositive, nominativ-appositive, nummerierend-appositive, (eigentliche) quantitative, additive, restriktive, parenthetische, juxtapositionale, präzisierende SyntRel**; in the field of coordinative relations: **koordinierende, sentential-koordinierende, koordinierend-konjunktionale, kommunikativ-koordinierende SyntRel**, and in the field of auxiliary relations: **passiv-analytische, auxiliare, quantitativ-auxiliare, korrelative, proleptische, elliptische SyntRel**.

5. Conclusions

At this stage of development the system of German dependency syntax proposed here comprises 58 surface syntactic relations; 18 of them have definitions that are

[9] cf. Apresjan et al. (2010:25–43).

similar to those of their Russian counterparts and differ from them just in some details. Seven relations have the same definitions as their Russian counterparts except for the concrete German lexemes that are part of the definitions. For another 30 relations identical definitions as for the corresponding Russian relations can be used. Only three German relations do not have a Russian counterpart, namely the *determinative Relation*, the *präfigurative* and the *reflexiv-analytische Relation*. Future work will show whether maybe one or another additional relation will be needed in German when a larger corpus with more different German constructions will have been analyzed.

Acknowledgements

Many thanks to three anonymous reviewers for their comments on the first version of this paper.

References

1. *Apresjan Ju. D.* et al. (1989). Lingvisticheskoe obespechenie sistemi ÈTAP-2 [Linguistic foundation of the system ÈTAP-2]. Moscow.
2. *Apresjan Ju. D.* et al. (1992). Lingvisticheskij protsessor dlja slozhnych informatsionnyh sistem [A linguistic processor for complex information systems]. Moscow.
3. *Apresjan Ju. D.* et al. (2003). ÈTAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the Meaning <=> Text Theory. Conference Proceedings of MTT 2003. Paris, 2003, pp. 279–288, available at http://proling.iitp.ru/publications/.
4. *Apresjan Ju. D., Boguslavskij I. M., Iomdin L. L., Sannikov V. Z.* (2010). Teoreticheskie problemy russkogo sintaksisa. Vzaimodejstvie grammatiki i slovarja [Theoretical problems of Russian Syntax. Interaction of grammar and dictionary]. Moscow.
5. *Boguslavsky I.* et al. (2011). Rule-Based Dependency Parser Refined by Empirical and Corpus Statistics. Conference Proceedings of International Conference on Dependency Linguistics, DepLing 2011. Barcelona, 2011, pp. 318–327, available at: http://depling.org/proceedingsDepling2011.
6. *Engel U.* (2009). Syntax der deutschen Gegenwartssprache. Berlin.
7. *Helbig G., Buscha J.* (2001). Deutsche Grammatik. Berlin, München.
8. *Kunze J.* (1975). Abhängigkeitsgrammatik. Berlin.
9. *Mel'chuk I. A.* (1974). Opyt teorii lingvisticheskih modelej «Smysl <=> Tekst». Semantika, sintaksis [Towards a theory of «Meaning <=> Text» models. Semantics, syntax]. Moscow.
10. *Heringer H. J.* (1996). Deutsche Syntax. Dependentiell. Tübingen.
11. *Specht V.* (2005). An MTT-Based Parser for German, in Apresjan Ju.D. (ed.). Vostok — Zapad [East — West]. Moscow, 316–329.
12. *Zangenfeind R.* (2011). Transfer of Russian Actantial Syntactic Relations into German, in Boguslavsky I., Wanner L. (ed.): Meaning — Text Theory 2011. Barcelona, S. 306–315, available at: http://meaningtext.net/mtt2011/proceedings/.

EXPLORING CONTEXT CLUSTERING FOR TERM TRANSLATION

Zhila A. (alisa_zh@mail.ru),
Gelbukh A. (gelbukh@gelbukh.com)

Center for Computing Research, Instituto Politécnico Nacional, Mexico City, Mexico

Many tasks in natural language processing, such as machine translation, word sense disambiguation, word translation disambiguation, require analysis of contextual information. In case of supervised approaches this analysis is performed by human experts, which is very costly. Unsupervised approaches offer fully automatic methods to fulfill these tasks. Yet these methods are not robust, their results are very parameter-dependent and difficult to interpret. Context clustering is an unsupervised technique for analysis of context similarities. In this work we explore dependencies of context clustering results from various clustering parameters. We also explore suitability of the context clustering for word translation disambiguation by evaluating the clustering results against known classes that are classes of translation candidates.[1]

Key words: translation, translation candidates, clustering, unsupervised methods, parameter, word sense discrimination, context

1. Introduction

In natural language processing word sense disambiguation is the task of automatic assignment of a correct sense from a predetermined sense inventory to a polysemous word. It is tightly related to the task of machine translation, where a correct translation of a word or phrase must be chosen from a list of translation candidates. Recently, the task of selection of the best translation or several interchangeable (synonymous) translations for a given source word in context and a set of target candidates has become known as word translation disambiguation.

In the past decade approaches to bootstrap machine translation with preliminary word sense disambiguation or word sense translation were explored in [1, etc.]. These approaches are based on supervised WSD classifiers that require extensive training on a large manually tagged training corpus. They are resource-demanding and provide relatively little or no improvement at a high cost.

The task of word translation disambiguation was treated independently in [4, etc.] either with a supervised classifier or with huge annotated monolingual corpora.

[1] Доклад публикуется в сокращенной версии. Полная версия доступна на сайте конференции «Диалог».

Context clustering is an unsupervised approach to detection of similarities in contexts [6, 10]. Its results highly depend on parameters used for clustering. This approach was applied to word sense discrimination [9], which is mere distinguishing between different senses.

The suitability of unsupervised approaches to word translation has not been explored yet. Our hypothesis is that unsupervised context clustering along with word aligned parallel texts can serve for obtaining context characteristics that would allow correct selection of a translation candidate for a word in a context in unsupervised manner. In this work we explore the suitability of context clustering for word translation disambiguation by comparing clustering results for various parameter combinations and evaluating them against known translation classes. In particular, we explore several parameter combinations with values that were found to be the best for document and context clustering in [11, 6, 9, 3, 12]. For evaluation of clustering results we use translation equivalents that were obtained from word aligned parallel corpus.

2. Context Clustering

In the past decade the topic of unsupervised word sense discrimination, that is discrimination between different word usages *in context*, was actively investigated. The most known solution to this problem is clustering of contexts that contain a word in question, which is a particular application of document clustering [6, 10]. An extensive review of clustering as unsupervised classification of dataset elements into groups is provided in [2]. The clustering algorithms suitable for document clustering are described and analyzed in [11] and implemented in SenseClusters clustering tool [8]. We adopt it as a tool for our experiments.

However, results of context clustering highly depend on a variety of parameters. Here we give a brief overview of clustering parameters and techniques.

2.1. Features

To perform a clustering one has to choose features that would represent each element of a dataset. In the field of document and context clustering features are sequences of words. *Unigrams* are one-word features. Pairs of two consecutive words are called *bigrams*. The extended definition in [3] states that bigrams are pairs of words that occur in a given order within some distance (*window*) from each other. Unordered pairs of words within a given window are called *co-occurrences*. For contexts that contain a marked word as in the case of word sense discrimination, *target co-occurrences* are introduced.

Features that occur fewer times than a threshold (*frequency-cut parameter*) r cannot serve as a solid basis for context grouping and, hence, must be excluded from the feature list as well.

2.2. Order of context representation

The first-order representation represents a context as a vector only of those features that are directly present in the context. The second-order representation considers features that co-occur with the features from the initial context in other contexts.

2.3. Similarity measure

To evaluate similarity between contexts, a *similarity measure* must be introduced on the selected feature representation. If elements are represented as feature vectors, such similarity measures as distance or cosine can be used. Hence, contexts can be either represented in a *vector space*, where a vector corresponds to each document, or a *similarity matrix* can be constructed based on pairwise similarities between contexts.

2.4. Clustering criterion functions

The task of clustering is optimization of a clustering criterion function, which is a function from similarity measure. A review and comparison of criterion functions is presented in [6]. The authors show that two of the compared criterion functions (I_2 and H_2) steadily provide good results with most of the clustering algorithms.

2.5. Clustering techniques

A detailed description of clustering techniques and comparison of their application for document clustering can be found in [11]. Their research showed that the best clustering techniques for document clustering are bisecting k-means and refined agglomerative clustering with UPGMA (Unweighted Pair-Group Method with Arithmetic Mean).

2.6. Cluster stopping criteria

Existing clustering techniques imply that a number of clusters is known in advance, which is not true in many cases, especially for context clustering and word sense discrimination. For cases when a desired number of clusters is unknown, automatic cluster stopping criteria were suggested in [3]. *Gap* is based on gap statistics applied to within-cluster dispersion, *pk1*, *pk2*, and *pk3* ("*pk*" stands for "Predicting the number of clusters K") consider significance of the change of a cluster similarity function between sequential number of clusters.

2.7. Clustering evaluation

Ranking of clustering algorithm performance depends substantially on what clustering quality measures measure is used [11]. There are two basic approaches to clustering evaluation: internal and external[2]. Internal measures do not use any external knowledge about possible groupings of clustered elements. External quality measures compare clustering results to a set of external known classes of clustered elements.

We evaluate clustering results against the set of translation equivalents obtained from a parallel corpus (see Section 3). We chose the widespread external measures of *entropy* and *purity*. We adopt their definitions and formulas from [11]. The entropy looks at uncertainty of a class distribution via clusters, while the purity evaluates how good a class corresponds to one cluster. In brief, the lower is the entropy and the higher is the purity, the better.

3. Experimental Settings

This section describes the performed experiment. It is aimed at exploration of how well unsupervised context clustering forms clusters that would be appropriate for word translation disambiguation. For this experiment, first, we use SenseCluster toolkit to perform unsupervised context clustering on our dataset.

We base our experiment on the results obtained in [7, 9, 11, 13] and explore whether these clustering parameters are appropriate for translation detection. The main assumption is that translations are correlated with a sense of a translated word. Unsupervised word sense discrimination holds that similar senses will be grouped into one cluster. Consequently, the contexts in a cluster would correspond to one translation equivalent or to several synonymous translations. We use cluster stopping criteria to automatically detect the number of word senses of an investigated word. We then compare the resulting number of clusters to the number of senses provided by dictionaries, detecting the most appropriate cluster stopping criteria.

3.1. Dataset

We used sentence aligned English-Spanish Europarl parallel corpus from OPUS open corpus [12] to extract contexts for clustering and to detect translation equivalents.

For our purpose of exploring context clustering suitability for word translation disambiguation, an ambiguous word had to satisfy the following criteria:

- to have a number of instances in a chosen parallel corpus that would be sufficient for unsupervised clustering (we set it 1000, which is about 2 to 5 times more

[2] Please, note that this is different from criterion functions.

than senseval-2 datasets used in [9] and realistic enough to be extracted from a corpus);
- to have more than one candidate translation in the parallel part of a corpus.

The analysis of the above criteria was performed using OPUS word alignment database. Several words satisfied these criteria: *facility*, *post*, *language*. Due to limited space, we present results only for the word "facility".

As a context we used an extract of 7 consecutive sentences from the corpus, one of them containing the chosen word. We chose the seven sentence context size basing on the average lengths of contexts used in the reference works [7, 9, 11, 13]. At this step we extracted 1771 contexts for our dataset.

We deleted wrongly aligned contexts from our dataset. The cases when the word "facility" did not have a direct translation equivalent were tagged as NOTAG. We grouped low-frequency (from 1 to 6) translation equivalents with their synonyms considering their context usage. In the end, we obtained a dataset of 1429 contexts and 21 translation classes including NOTAG[3]. These translation classes serve for external evaluation of obtained clusters.

According to the monolingual dictionaries, which we consulted (Online Merriam-Webster, Oxford Concise Thesaurus, WordNet, and Larousse American Pocket), 4 to 5 senses for the word "facility" can be distinguished. We took these numbers as guidance for the minimum number of clusters. Therefore, any combination of parameter values that gave fewer than 4 clusters was discarded.

3.2. Clustering parameters

In this work we perform context clustering with SenseClusters [8]. It is a complete and freely available context clustering system that provides support for feature selection from large corpora, several different context representation schemes, various clustering algorithms, and evaluation of the discovered clusters.

Parameters with fixed values We set values of several parameters to be unchangeable for our experiments:
- the order of feature representation is set to *-o2*, which stands for the second order;
- the context are represented as feature vectors in *vector space*;
- window is set to 5;
- frequency-cut parameter *r* is set to 3.

We chose the second-order context representation since it is shown to be better for short contexts [5]. The vector space is preferred over the similarity matrix representation based on the work [9]. For the value of window parameter we took as a reference [7], where this parameter was set to 5. We set the value of the frequency-cut parameter *r* to 3 heuristically basing on [7].

[3] The dataset is available at *www.gelbukh.com/resources/word-translation-alignments*.

Parameters with varied values Since the total number of possible combinations of parameter values is very high, we analyzed only among those parameter values that are proved to be the best for document and context clustering in [11, 13]. The parameters with the values that we varied are:

- features for context representation: unigrams, bigrams, co-occurrences, target co-occurrences;
- clustering methods: direct k-means, repeated bisection, refined repeated bisection, agglomerative;
- criterion functions: I_2, H_2, and UPGMA;
- cluster stopping measures: gap, pk1, pk2, pk3.

The total number of experiments is 112.

4. Experimental results

The number of clusters that we obtained with various clustering parameter combinations varied from 1 to 6.

Cluster stopping measures. Table 1 shows the frequencies of each number of clusters for a cluster stopping measure.

Table 1

cl. num.	1	2	3	4	5	6
gap	24	0	4	0	0	0
pk1	11	10	3	1	3	0
pk2	0	8	10	3	4	3
pk3	0	12	9	6	1	0

As it follows from Table 3, cluster stopping measures *gap* and *pk1* provide the lowest number of clusters. Gap statistic measure gives no results that would be higher than the threshold of 4 clusters. Pk1 measure gives acceptable results only in 4 cases, which is 3.5 % of all cases.

The fractions of experiments for each cluster number from the total number of experiments are shown in Table 2.

Table 2

cl. num.	1	2	3	4	5	6
fraction, %	31.2	26.8	23.2	9.0	7.1	2.7

Of the total number of experiments 50 % were for cluster numbers 2 and 3, and only 18.8 % (21 of a total of 112 combinations) passed the threshold of 4 clusters.

An assumption that the word "facility" may have only 2 to 3 *"real"* or well distinguishable senses does not seem to be probable. If one refers to the dictionary definitions of "facility", its meanings can hardly be grouped into a number of independent and semantically poorly related senses less than 4.

Therefore, we interpret the steadily low number of clusters for cluster stopping criteria gap and pk1 as an inherent quality of these criteria. *Pk2* and *pk3* measures give acceptable results in 36% and 25% of their usage cases respectively.

Context features, entropy and purity. The values of the parameter, entropy, and purity for the experiments, for which the number of clusters resulted to be at least 4, are presented in Table 3. We remind that the number of senses for "facility" given by monolingual dictionaries is 4 to 5.

No experiments with bigram and target co-occurrences features gave the number of clusters more than or equal to 4. Therefore, the results of all experiments with these features were discarded and are not shown in Table 3.

Notations of Table 3 are as follows: *clmeth* stands for "clustering method", *crfun* is "criterion function", *clstop* designates "cluster stopping measure", *cl #* is the resulting number of clusters for a given combination of parameters, *E* and *P* are entropy and purity respectively. In the table we use abbreviations: *agglo* for agglomerative clustering method, *direct* for direct k-means, *rb* for repeated bisection, *rbr* for refined repeated bisection. These notations and abbreviations are used through the rest of the paper. Other notations are explained in Section 2 and in Section 3.2.

Table 3

clmeth	crfun	clstop	cl #	E	P	clmeth	crfun	clstop	cl #	E	P
Co-occurrences						Unigrams					
agglo	upgma	pk2	6	80.6	25.5	agglo	upgma	pk2	6	84.1	24.2
direct	h2	pk1	4	80.4	25.6	direct	i2	pk2	6	**74.8**	26.9
direct	h2	pk3	4	80.4	25.6	rb	h2	pk1	5	75.2	**28.3**
direct	i2	pk2	5	80.2	25.5	rb	h2	pk2	4	76.2	27.6
direct	i2	pk3	4	80.4	25.6	rb	h2	pk3	4	76.2	27.6
rb	h2	pk1	5	80.7	25.0	rb	i2	pk2	5	75.6	27.8
rb	h2	pk2	4	81.0	25.0	rbr	h2	pk1	5	75.2	**28.3**
rb	h2	pk3	4	81.0	25.0	rbr	h2	pk2	4	76.2	27.6
rb	i2	pk3	5	80.7	25.0	rbr	h2	pk3	4	76.2	27.6
rbr	h2	pk3	4	80.4	25.6	rbr	i2	pk2	5	75.3	**28.3**
rbr	i2	pk2	5	80.2	25.5						

In table 3 the best entropy and purity values are in bold. They all occurred for the parameter combinations with unigram features. Generally, the entropy and the purity for unigrams is about 5% and 12% better respectively than for co-occurrences. Yet comparison of the entropy and purity values to those obtained in [11, 13] is hindered by the dependence of entropy and purity on the number of classes.

Our consideration is that the entropy and purity measures as they are described in Section 2.7 might be inappropriate for our task. These measures were intent

to evaluate word sense discrimination results, when it is assumed that each cluster corresponds to a sense and it is expected (or manually set) that the number of clusters would be more or less similar the number of senses. In our case it is acceptable for more than one class to be clustered together or for elements of one class to be distributed between several clusters as far as it is interpretable from the translational point of view.

Number of clusters. To check how cluster number influences the entropy and purity, we perform an experiment with a fixed number of clusters manually set to 21, which is the number of our translation classes. In this experiment we used a clustering parameter combination that gave the highest purity. The results are shown in Table 4.

Table 4

clmeth	crfun	clstop	cl #	E	P
Unigrams with fixed number of clusters					
rb	h2	n/a	21	67.2	32.7

As it can be seen, the more than fourfold increase of the cluster number from 5 to 21 improves the values of entropy and purity only 10.6 % and 15.5 % respectively.

Illustration of clustering results. For illustration of clustering results we chose two of the best experiment cases: one for the co-occurrence feature and another for the unigram feature. Table 5 shows the distribution of translation classes through resulting context clusters for two parameter combinations.

Table 5

clD	cl. size	disp	equi	Faci	instal	*NO*	serv	sist	posib	meca	instit	capac	centro	ayud	medio	crédi	planta	fond	medida	infra	plan	central
	Co-occurrences, Agglomerative Clustering, UPGMA, Pk2, Entropy = 0.806, P= 0.255																					
0	989	42	22	121	196	94	96	19	95	82	16	25	18	19	55	18	16	20	8	14	4	9
1	3	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
2	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	350	3	6	24	142	40	40	3	14	6	10	4	9	0	20	1	6	1	5	2	0	14
4	11	0	0	5	1	1	0	0	2	0	0	0	1	0	0	0	0	0	0	1	0	0
5	75	18	1	1	4	3	5	19	4	8	2	0	1	2	1	1	0	5	0	0	0	0
	Unigrams, Repeated Bisections, H2, Pk1, E= 0.758, P=0.276																					
0	213	27	1	24	8	17	16	6	24	26	1	9	3	7	12	9	1	15	1	3	2	1
1	156	1	0	1	112	10	3	0	1	1	0	0	1	0	5	0	8	0	2	0	0	11
2	282	0	13	42	83	26	35	3	22	4	6	4	8	3	13	1	3	0	2	9	0	5
3	307	32	1	29	24	25	11	25	32	57	4	9	6	7	18	7	3	10	4	0	2	1
4	471	3	14	56	116	61	77	7	36	8	17	7	11	4	28	4	7	1	4	5	0	5

In the beginning of Section 3 we explained the assumption that unsupervised context clustering would be suitable for word translation selection if a cluster

corresponded to one or more entire translation classes, which is the case of synonymy between translations, or if a translation class was distributed between *some* clusters, which is the case of preserved cross-lingual homonymy. For these cases a clustering solution would tend to obtain "neat" groupings of classes per clusters and intuitively we can predict that a class-cluster distribution table would have more zeros than non-zeros. However, nearly all cells have non-zero values. It means that a context corresponding to any translation equivalent can be found in any cluster. This violates our initial assumption about context clustering suitability for word translation selection.

5. Conclusions and future work

In this work we perform comparison of various clustering parameter combinations and explored suitability of context clustering application to unsupervised word translation.

The number of clusters more than the threshold of 4 occurred only for 18.8% of the experiments. Numbers of 2 and 3 clusters were detected in 50% of cases. At this step these results could not be interpreted from the semantic point of view and were discarded. However, formal analysis of semantic similarity of senses through an ontology or semantic hierarchy can give new perspective on these numbers.

We detected that cluster stopping measures *gap* provides very low numbers of clusters that cannot be interpret from the semantic point of view. The numbers of clusters that correspond to the semantic assumption of the number of word senses can be achieved in most cases with *pk2* and *pk3* cluster stopping measure. Also *pk1* cluster stopping measure should not be completely discarded since it provided 19% of all acceptable results.

We were not able to detect acceptable results for *bigram* and *target co-occurrence* features. It might be explained by inappropriate window size and data sparseness that in our experiments was not handled through singular value decomposition. Further experiments with singular value decomposition and varying window size are necessary.

The evaluation of results through entropy and purity gives us the numbers that are not easily interpreted in the task of word translation when the number of classes is much higher than the number of clusters. Hence, we will work on development of different quality measure that would be more adequate for our goals.

References

1. *Carpuat M., Wu D.*, Improving statistical machine translation using word sense disambiguation, Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007), 2007, pp. 61–72.
2. *Jain A. K., Murty M. N., Patrick J. Flynn P. J.* (1999), Data Clustering: A Review. ACM Computing Syrveys, vol. 21, pp. 264–323.

3. *Kulkarni A., Pedersen, T.* (2006), Unsupervised Context Discrimination and Automatic Cluster Stopping, MS Thesis, University of Minnesota Supercomputing Institute Research Report UMSI 2006/90.
4. *Marsi E., Lynum A., Bungum L., Gambäck B.*, Word Translation Disambiguation without Parallel Texts. Proc. International Workshop on Using Linguistic Information for Hybrid Machine Translation, Barcelona, Spain, 2011.
5. *Pedersen T.* (2008), Computational Approaches to Measuring the Similarity of Short Contexts: A Review of Applications and Methods, University of Minnesota Supercomputing Institute Research Report UMSI 2010/118.
6. *Pedersen T., Bruce R.*, Distinguishing word senses in untagged text, Proc. of the Second Conference on Empirical Methods in Natural Language Processing, Providence, RI, 1997, pp. 197–207.
7. *Purandare A.* (2004), Unsupervised Word Sense Discrimination By Clustering Similar Contexts. MS Thesis. University of Minnesota.
8. *Purandare A., Pedersen T.*, SenseClusters — Finding Clusters that Represent Word Senses. Proc. of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-04), 2004, pp. 26–29.
9. *Purandare A., Pedersen T.*, Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces, HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004), 2004, pp. 41–48.
10. *Schütze, H.* (1998), Automatic Word Sense Discrimination. Journal of Computational Linguistics, vol. 24(1), pp. 97–123.
11. *Steinbach M., Karypis G., Kumar V.* (2000), A comparison of document clustering techniques, University of Minnesota, Technical Report 00–034.
12. *Tiedemann J.* (2009), News from OPUS — A Collection of Multilingual Parallel Corpora with Tools and Interfaces, Recent Advances in Natural Language Processing, vol. V, pp. 237–248.
13. *Zhao Y., Karypis G.* (2001), Criterion Functions for Document Clustering: Experiments and Analysis, University of Minnesota, Department of Computer Science Technical Report 01–040.

A UNIFIED ANALYSIS OF CLITIC CLUSTERS IN WORLD'S LANGUAGES[1]

Zimmerling A. (meinmat@yahoo.com)

Moscow State University for the Humanities, MSUH

The paper proposes a unified analysis of complex syntactic objects defined as *clusters*. A cluster is by definition a string of elements {a,b,c…n} which can function without combining with each other but are arranged in a rigid order when they assume a contact position, so that for each pair (a, b) the linear order a > b, i. e. 'a precedes b' is fixed. Elements conforming to this definition are called *clusterizing*. Rules ordering elements in clusters are called *Template Rules*. In the first section I analyze Template Rules as empiric generalizations made on text corpora representing the normative usage of world's languages from the class of languages with clusterizing clause-level elements. In the final section I analyze Template Rules as linearization algorithms. The general conclusion is that clusters ordered by Template Rules are normally non-homogenous regarding their morpho-syntactic and prosodic values. I furthermore argue that a theory of clusters can be build with little or no resource to the prosody of the clitic elements.[2]

Key words: clusterization, clitics, syntax, linearization algorithms, template rules

Clusters of sentence elements are attested in a large class of languages, cf. [Anderson 1993]. Almost all clusterizing elements attested in natural languages are *clitics,* i. e. prosodically deficient expressions linearized by specific rules presumably not affecting non-clitic elements, cf. Zwicky (1977). At the same time, not all clitics clusterize. In recent versions of the Minimalist Program, standard non-clitic elements (except for the *pro*-forms) are treated as maximal projections (XP), while clitics are treated as heads (°X), cf. [Franks 2008] or as so-called left-branching elements (XP/°X) i. e. reduced groups, cf. [Bošković 2001]. Consequently, discussion of clusters often starts from stating the researcher's credo — whether (s)he believes that clitics are °X-s or XP/°X-s, and goes on global architecture of language. As a result, analysis of clusters in the framework-based research relies on issues not directly related to the surface order of elements — whether syntax and morphology are one module of grammar, on the linguist's wish to go in for prosody of the clusterizing elements etc. If a framework-based author adopts a definition of a clitic as an element 'lacking stress' and comes across clusters consisting of both unstressed and stressed elements, (s)he will likely conclude that these formations are not clusters. In short, current linguistic theories

[1] The paper was written with support of the Russian foundation for the humanities, project RFH 11-04-00282a. I am grateful to the anonymous reviewer for the valuable comments.

[2] Доклад публикуется в сокращенном варианте. Полная версия доступна на сайте конференции «Диалог»

capitalize the idea that clusterizing elements are clitics, explain fixed combinations of adjacent elements with special features postulated for clitics without postulating special features for clusters and impose conditions that clusterizing elements should be homogeneous. This condition may be called Homogeneity Condition:

(i) Homonegeity Condition, HC.
Clitics clusters must consist of clitics sharing all relevant prosodic and morpho-syntactic features typical of clitics in this language.

The main drawback of the approach outlined above is that no one has proved yet that clusterizing and non-clusterizing clitics have the same properties. Moreover, it is unclear whether these two sets are intersecting or not. Let us define non-clusterizing clitics formally.

(ii) A non-clusterizing clitic is a phrase-level or clause-level element that does not clusterize with other elements of the same level.

According to the proposed definition a single clitic *c* is considered clusterizing if it can be substituted with a cluster [$_{CLP}$ abc] containing the element *c* in other phrases with the identical fragment of syntactic structure. E.g., if a language has both reflexive and auxiliary clitics and they clusterize as [$_{CLP}$ CL_{Refl}, CL_{Aux}], a single auxiliary clitic is considered clusterizing if it appears in a sentence without a reflexive verb. Similarly, if a language only has 1–2 p. auxiliaries which clusterize with reflexives but lacks 3 p. auxiliary clitics, a single reflexive clitic is considered clusterizing if used in a sentence with a verb in the 3 p. If a language has a question marker which makes up clusters with pronouns in yes-no questions, its absence in non-interrogative clauses using the same set of pronominal clitics does not falsify the hypothesis that this marker and a pronominal clitic clusterize if they both are realized overtly in the same clause. I conclude that the optional character of clitics, selective restrictions imposed on their combinations and correlations between subsets of clitics and types of speech acts are not sufficient for denying clusterizing capacities of a clitic. Meanwhile, there are indisputable examples of non-clusterizing clitics: these are elements that either do not occur in adjacent position to any other clitics of the same type[3] and phrase-level or do not form fixed sequences with them. E.g., Serbo-Croatian (S-C) has a set of accusative enclitics which attach to prepositions, cf. (1a): they do not combine with any other enclitics and are phrase-level (PP-level) elements that do not climb i. e. cannot be extracted to a higher domain, cf. the ill-formedness of (1b).

(1) S-C a. Da =*je* Ivan računao [$_{PP}$ na=*me*].
'that Ivan counted on me.'
b. *[$_{CP}$ da=***me**$_{i}$ =*je* Ivan računao [$_{PP}$ na ___$_{i}$]].

[3] This stipulation is necessary if we want to separate possible proclitic + enclitic sequences, with a proclitic serving as a host element for the enclitic and a single enclitic attached to it, cf. example (1b), from clusters consisting of two or more proclitics/two or more enclitics.

S-C also has segmentally identical accusative enclitics *me*$_{\text{ACC.1SG}}$, *te*$_{\text{ACC.2SG}}$ which are clause-level elements encoding predicate arguments. Clause-level accusative enclitics in S-C clusterize with other enclitics and climb[4]. The non-clusterizing *me*$_{\text{ACC.1SG}}$ and the clusterizing *me*$_{\text{ACC.1SG}}$ have the same morphological marking but different syntactic properties. If syntax of S-C *me*$_{\text{ACC.1SG}}$ crucially depends on its domain — clause vs phrase, it seems better to analyze the clause-level clusterizing S-C clitic *me*$_1$ and the phrase-level non-clusterizing S-C clitic *me*$_2$ as homonyms but not as one and the same underlying syntactic element.

Another challenge to the hypothesis that the sets of clusterizing and non-clusterizing clitics intersect comes from the distribution of particles. In a number of languages indeclinable particle clitics convey different meanings in configurations where they clusterize and where they do not. E.g., in Old Novgorod Russian (O-N) the enclitic particle *=že* was non-clusterizing if conveyed an identifying meaning, cf. *totъ=že* 'the same', or an additive meaning 'with X and with Y' but clusterizing in all other meanings. It is possible to associate the identifying and additive meanings of *=že*$_2$ with non-sentential uses, where it does not pertain to the clausal predicate, and to separate them from the clusterizing particle *=že*$_1$. The clause-level *že*$_1$ clusterizes with other clause-level clitics, since it is a clause-level operator itself, while the phrase-level *že*$_2$ does not clusterize, since it is an NP-level element which serves as a means of highlighting particular discourse referents. Here different syntactic status of the NP-level non-clusterizing clitic *že*$_2$ vs clause-level clusterizing clitic *že*$_1$ corresponds to differences in semantics, while their prosodic features do not change.

There are languages where clusterizing clitics and their non-clusterizing counterparts differ in the prosodic type. E.g., in Slovak (Svk) the particle *=už*$_1$ 'already', 'still' is clusterizing if it is an enclitic, cf. (2a) and non-clusterizing if it is a proclitic *už*$_2$*=*, cf. (2b). The clusterizing *=už*$_1$ cannot be clause-initial and is invariably placed after clusterizing pronouns like *=im* in (2a), while the non-clusterizing *už*$_2$*=* can and occasionally precedes enclitic pronouns as in (2b). If one treats *=už*$_1$ and *už*$_2$*=* as one and the same underlying syntactic element and not as a pair of homonyms, one will arrive at the wrong conclusion that the enclitic particle *=už*$_1$ is not part of the Svk clitic cluster, since the relative order of { *už*$_1$ } ∪ {*už*$_2$} and Svk pronominal enclitics is apparently invertible. This bug is eliminated if we separate *už*$_1$ and *už*$_2$ as elements with different categorial features and do not treat the proclitic + enclitic sequences like #**už** *=ju* in (2b) as clusters.

(2) Svk a. Ale [$_{\text{PP}}$ v hľbke duše] *sa* *=im* ***=už*** tešil.
'But deep down he **still** found comfort in the hope <that P>'
b. #**Už** *=ju*$_{\text{i}}$ aj začinajú [$_{\text{IP}}$ robiť __$_{\text{i}}$].
'One has **already** started building *it* <the road>'.

Finally, in a number of languages clusterizing clitics have counterparts in non-clitic expressions with the same segmental structure. E.g. Czech clusterizing pronouns

[4] Cf. [Ćavar, Wilder 1999: 447–451] for the discussion of climbing out of infinitive clauses in S.C.

in the 3 Sg. feminine are homonymic to stressed pronouns in the same case, number and gender. In Warumungu this correlation is extended to all personal pronouns: they are described as clitics when used clause-internally but as stressed non-clitic forms in the clause-initial position [Mushin, Simpson 2008: 585]. Old Icelandic had clause-internal deictic particles *nú* 'now' and *þá* 'then' which clusterized with pronouns and adverbs [Zimmerling 2002: 368]. The clause-internal clusterizing enclitics $nú_1$ and $þá_1$ had non-clusterizing clause-initial counterparts $nú_2$ and $þá_2$ that likely were non-clitic elements bearing the sentence accent. The same problem arises with Old Russian 1–2 p. forms of the BE-auxiliary in the present tense indicative — they are distinguished from their non-clitic counterparts merely on the basis of their syntactic position.

Similar facts do not prove that clusterizing clitics and their non-clusterizing counterparts are homonyms but they compromise the idea that clusters are just occasional sequences of prosodically homogeneous non-clusterizing clitics. I am therefore giving up the HC condition and conclude that the parameters '± clitic' and '± clusterizing' are independent: each slot in the binary matrix on Fig. 1 may have a non-empty filling.

	A.'+ Clitic'.	**B.'-Clitic'.**
1. '- Cluster'.	+	+
2. '+ Cluster'.	+	CLITICOIDS

Fig. 1. Clitics and clitic clusters in world's languages

The existence of A1 and B2 is an observable fact. The existence of A2 can be proved formally if a language also has a non-identical set of non-clusterizing clitics and all clitics involved in clusters met all the conditions met by non-clusterizing clitics in this language. The existence of B2 can be proved if some or all clusterizing elements do not conform to the conditions met by clitics in this language. If they combine in a cluster with standard clitics, the value '+clitic' is not the most general feature of the clusterizing elements. I am tagging clusters consisting of elements which violate the HC as 'cliticoids'. Informally, cliticoids are a set of clusterizing elements which share some common value overlapping but not directly matching with the value '+clitic', e. g. a value like '+function word', or '+discourse-activated element' etc. What a linguist usually does not know beforehand is whether B2 and A2 overlap or A2 is a proper subset of the B2. Empirically-oriented research has shown that world's languages usually allow clusters consisting of elements representing two, three or more different kinds of sentence categories, e. g. Agreement Markers, Case Markers, Reflexives, Discourse Operators. Therefore, the most general value assigned to all cliticoids is less likely to be found in the set of morphosyntactic features and more likely to be found in the set of communicative or prosodic features.

I claim that the distribution of clusters provides a key for retrieving information about syntactic domains. I also argue that the list of diagnostic features of clusters can be construed without assuming that all clusterizing elements are clitics. Moreover,

not all sequences of adjacent clitics are clusters even if the order of the elements is fixed and they obey HC.

A salient counterexample is furnished by cases where adjacent clitics *a*, *b* belong to different categories X, Y and the fixed order a > b is straightforwardly derived from the fixed order X > Y. If X и Y are sentence categories that are placed in a fixed order [X], [Y], the clitic *a* can attach to X° only from the right (X°=a)[5], the clitic b can attach to Y° only from the left (b=Y°), the fixed order a > b is generated automatically: [$_X$ X°=a] [$_Y$ b=Y°]. To eliminate such sequences and to keep them apart from clusters where all elements belong to one and the same syntactic domain [$_X$ X°=ab] or [$_Y$ ab=Y°], one needs a principle like (iii).

(iii) The Independent Templatic Principle,ITP.
If elements from the set {a,b,c…n} constitute a cluster, the order a > b for the each pair of adjacent elements (a,b) is predicted by an independent Template Rule which cannot be derived from the rules predicting the order of any pair of sentence categories X, Y, X$\notin$ {a,b,c…n}, Y $\notin$ {a,b,c…n}.

It is reasonable to impose a further condition on clusters and specify that the clusterizing elements should represent one syntactic domain.

(iv) The One Domain Principle, ODP.
If clitics from the set {a,b,c…n} are ordered by a Template Rule, predicting the order a > b for each pair of adjacent elements (a,b), the cluster /a,b,c…n/ cannot be divided by any categorial brackets dividing the domains of any two syntactic categories X and Y, so that one part of the cluster would belong to X, and the other one to Y: *[$_X$ …ab][$_Y$ cd…].

We are unable to prove ODP in the strong sense i. e. to prove that adjacent clitics *a, b* capable of taking more two or more positions in clause, *cannot* be placed in a fixed order a > b if they belong to different domains X and Y. ODP probably can be proven in the weak sense i. e. for each language which has clusters of clause-level elements /a,b,c…n/ it can be demonstrated that the clusterizing elements belong to one domain (not two or more).

Sequences of clitics belonging to different domains resemble clusters but are not recognized as such according to ODP. Such sequences arise in languages where clitics of the matrix clause attach to the right edge of the embedded clause. This is attested in Cavineña which has phrase-final complementizers like *tibu* ‘because’ [Guillaume 2008: 84]. In (3) the enclitic *=tibu* belonging to the embedded clause with the verb *jipe-kware* ‘approach’, ends up to the left of the cluster *=pa=tu* belonging to the main clause predicate *pude-da* ‘(be) red’.

[5] The symbol ‘=’ is used here as a marker of cliticity, not as an equals sign. The notation X=*a* reads ‘clitic =a is attached to category X’.

(3) Cav [$_{CP}$ *Tu-ra*=kamadya ijeti jipe-kware=*tibu*]# **=*pa* =*tu*** pude-da.
'Since he <the flycatcher bird> is the only one who had approached the sun in the old days, he is red/brown.'

The ODP condition must be modified for languages that allow climbing of clitics. An example of climbing is given above in (2b). A more complicated case is shown in (4).

(4) Svk [$_{TP}$ Aj ty =*by* =*si* =*sa*$_i$ mal [$_{IP}$ zodvihnúť t$_i$] z tej postele!].
' It were better if you just arose from the sack!'

This condition is shown schematically in (4'), where the superscripts $b^1c_i^2$ refer to heads V1 and V2 and the subscript ($_i$) refers to the base-generation position of the moved ('climbed') clitic.

(4') [$_{TP}$ …a^1 b^{1}c$_i^2$ °V1 [$_{IP}$ … °V2 t$_i$]].

The well-formedness of structures like (4') indicates that ODP must be complemented by a condition on embedding and climbing.

(v) The Condition on Embedded Domains, CED.
A cluster defined on the set {a,b,c…n}, where for each pair of adjacent elements (a, b) the fixed order a > b can be established, and elements a,b,c,d are generated in different clauses of a multi-predicate complex [$_X$ abc$_i$d$_j$ [$_Y$ t$_i$ t$_j$…]], syntactically belongs to the hierarchically highest clause of this complex, if the Template Rule has slots {{$_C$ …}, {$_D$ …}} for the clitics c, d extracted from the embedded clause [$_Y$ …].

If the CED condition is met, structures with climbing like (4') do not violate the ODP.

Some sequences of adjacent clitics belonging to different domains and violating the ODP arise not at the clause junction as in (3) but in structures where clause-level clitics adjoin to the right edge of the preceding constituent which can include phrase-level clitics. Cf. examples like (5), where the bound clitic =*to* is part of the relative marker *o-že=to* and does not belong to the same level of representation as the free clusterizing clause level clitics =*mi*$_{DAT.1SG}$=*esi*$_{AUX.2SG}$.

(5) O-N [$_{TP}$ [$_{CompP}$ o-že=*to*] =*mi* =*esi* povedalъ].
'Since you have told me'.

In order to separate sequences with adjacent clitics of a different level from clusters consisting of elements of the same level, we add the condition (vi).

(vi) The One-Level Condition, OLC.
The Clusterizing elements must belong to one and the same level of representation. Phrase-level clitics do not clusterize with clause-level clitics.

The separation of adjacent clitics from clusters is a difficult task but not a conceptual problem. If a linguist has a sufficient corpus, the OLC usually makes it possible to verify or falsify the hypothesis that the *k* is a clusterizing element i. e. an element placed by the same Template Rule that predicts the order of clitics a, b, c, if a,b,c are elements of the same level. E.g., Warlpiri shows sequences like [=*ku*=*ju*] + [=*ka*=*rna*], as in (6), where =*ka* and =*rna* are clusterizing clause-level clitics. Yet the first two elements scarcely belong to the clausal domain: the element ju_2 is topic marker[6] and is adjoined only to nouns and stressed pronouns, cf. *nyuntu* in (6). The marker =*ku* is glossed as 'Dative' but it is again only possible on NPs, cf. the example (7), where =*ku* is supported by the predicate-level dative clitic =*rla* which is placed in the cluster after nominative and accusative enclitics, according to the Warlpiri Template Rule.

(6) Warl [$_{NP}$ Nuyntu=***ku*=*ju***$_2$]=*ka* =*rna* wiri nyina [$_{NP}$ ngaju=ju_2].
'I am bigger than you'.

(7) Warl [$_{NP}$ Kurdu-ngku]=*ka* =ju_1 ***=rla*** [$_{NP}$ ngaju=***ku***] paka-rni.
'The child is striking at me'.

Both =*ku* and the topical =ju_2 in (6) appear to be NP-level clitics which can immediately precede the clusterizing clause-level clitics =$ka_{AUX.PRS}$=$rna_{1SG.NOM}$ due to a special parameter setting for Warlpiri: this language allows the placement of clause-level clitics after the first spelled-out clausal constituent. If ITP, ODP, SED and OLC are instrumental in all languages or at least in a class of languages with clusterizing clause-level clitics, it is not necessary to learn all languages from this class in detail in order to separate adjacent clitic sequences from clusters consisting of syntactic elements of the same level and belonging to one and the same domain.

The difficulties with separation of adjacent clitics from clusters are salient in sentences like (3), (5), (6), (7) where clitics belonging to clusters are prosodically homogeneous with adjacent elements of a different phrase-level. Clusterizing elements of the same phrase-level belonging to one and the same syntactic domain often violate HC: some of them have non-clitic counterparts with the same segmental structure and morphological marking, cf. O-N clitic =$esi_{BE.AUX.2Sg}$~ O-N $ESI_{BE.AUX.2Sg}$, others have counterparts in the set of non-clusterizing clitics (cf. S-C clause-level =me_1 and phrase-level =me_2), or counterparts differing in the prosodic type, cf. Svk clusterizing enclitic =$už_1$ with the non-clusterizing proclitic =$už_2$. Such deviating features are not characteristic of *all* clusterizing elements in the corresponding languages. There are other examples showing the prosodic and morphologic heterogeneity of clusters consisting of syntactic elements of the same level and belonging to the same domain. E.g. clusters of clause-level clitics in Sanskrit alternated stressed and unstressed elements. Template Rules operate on clusterizing clause-level elements with a different

[6] At the synchronic level Warl. =*ju* is described as a pair of homonyms: the clusterizing pronominal 1 p. accusative clitic =*ju*, cf. [Nash 1986: 56, 59]. We mark the homonyms =ju_1 и =ju_2 with subscripts.

number of syllables and assemble clusters consisting of both monosyllabic and polysyllabic elements. Finally, in many languages the internal order in the cluster violates the so called Categorial Principle [Kosta, Zimmerling 2012]. This principle predicts that all elements are grouped in blocks according to their taxonomic category, e. g. Particle, Pronoun or Auxiliary, so that the blocks do not intersect in the Template: strings consisting of two or more particles cannot be intervened by any auxiliaries and vice versa. This condition is shown in (vii): the capital letters A, B, C are for different categories of clusterizing clitics, the lowercase letters with indexes a^1, b^1, c^n are for particular clitics representing categories A, B, C.

(vii) The Categorial Principle: the clusterizing elements are grouped according to the taxonomic morphosyntactic category. The order of non-intersecting blocks in a clitic template embodies some hierarchy of sentence categories.
$[_{\text{CliticPhrase}}\ [_{\text{A}}\ a^1, a^2..a^n]\ [_{\text{B}}\ b^1, b^2\ldots\ b^n]\ [_{\text{C}}\ c^1\ , c^2\ldots c^n]]$.

The condition in (vii) is preserved in Template Rules of several languages including O-N where all clusterizing particles precede all clusterizing pronouns, and all clusterizing pronouns precede all clusterizing auxiliaries. It is severely violated in Tagalog where monosyllabic particles intervene in sequences of two or more pronouns. It is mildly violated in Slovak where clusterizing auxiliaries and pronouns are placed between two slots for particles: the leftmost slot is filled by the conditional particle *=by*, the rightmost slots are filled by the spatial particles *=tu, =tam* and discourse particle *=už*.[7]

Particle1	**Auxiliaries 1–2 p.**	**Pronouns**				**Particle2**	
		Reflexive	**Dative**	**Accusative**	**Nom.-Acc. Sg. n.**		
by	sem, $si_{\text{2SG.PRS}}$, sme, ste	sa, $si_{\text{REFL.DAT}}$	mi, ti, mu, jej, nám, vám, im	ma, ťa, ho, ju, nás, vás, ich	to	tu, tam	už
1	2	3	4	5	6	7	8

Fig. 2. Slovak Template Rule

Fig. 2 shows that 8 slots of the Template host 25 clusterizing elements representing 3 categories — Auxiliary, Pronoun and Particle. The slots 1,6,7,8 are filled with a single element each, while slots 2,3,4,5 are filled with sets of elements which share at least two properties — a) they do not occur simultaneously; b) they are ordered in the same way respective all other elements filling all other Template slots. This distributional restriction is a general feature of all Template Rules. At the same time, the order of slots is *idiosyncratic* [O'Connor 2002]. The last term has a twofold load.

[7] Note that Slovak has two homographic and homophonous clusterizing clitics taking different slots in the template: the auxiliary clitic $=si_{\text{2SG.PRS}}$ and the dative reflexive clitic $=si_{\text{REFL.DAT}}$.

For the first, it states that the order of slots is *language-specific*, not universal. For the second, it prompts that the order of slots is not derivable from any other rules and principles except those which exclusively pertain to the clusterizing elements, cf. the condition (iv) above, introduced in this paper as an autonomous principle, Independent Template Principle, ITP.

(viii) The Template-and-Slots-Principle, TSP.

1. The number of slots in a Template equals the number of non-overlapping subsets A,B,C..N, so that each subset consists of elements $\{a_1, a_2 \ldots a_n\}$ $\{b_1, b_2 \ldots b_n\}$, which stand in a complementary distribution with each other and are ordered uniformly respective any other elements from other subsets.
2. The order of slots is idiosyncratic and cannot be derived from any other principles than ITP.

Let us now treat Template Rules not only as generalizations based on some text corpora but also as dynamic systems whose main function is to predict the order of any two elements $< x, y >$ from the strictly bounded and relatively small set of the clusterizing elements {a,b,c…n}. There are two scenarios. If *x, y* take the same slot, they stand a in complementary distribution and do not show up simultaneously. If *x, y* take different slots and may show up simultaneously[8], there are ordered in the only possible way x > y in all cases where they assume a contact position. If the clusterizing elements take a distant position and do not form a contact sequences, this requirement does not hold. That is, if *a°, b°, c°* are elements belonging to the cluster [$_{\text{CliticP}}$ <a°, b°, c°… n°>], sequences like *X=*b°* (2)=*c°* (3)= *a°* (1) are excluded if all these clitics are attached to one and the same sentence category X, but if *b°, c°* attach to X, while *a°* attaches to Y, sequences like X=*b°* (2)=*c°* (3) Y= *a°* (1) are possible. Contrary to [Bošković 2001:21], disjoint placement of clusterizing clitics does not by itself defy the existence of clusters if the TSP is not violated and configurations with cluster splitting can be proved to be derived from configurations without splitting. Rules triggering cluster splitting are called 'Barrier Rules' in the tradition based on [Zaliznjak 1993: 288][9]. Empirically-oriented research has shown that most if not all previously described word order systems with clusterizing clitics have Barriers. If this fact is not incidental, it hints that Barriers serve as a means preserving syntactic independence of the clusterizing elements, since they enlarge the list of positions available for them.

If one interprets the internal ordering in terms of *blocks*, i. e. sets of slots sharing the same taxonomic category (i. e. Auxiliary vs Particle vs Pronoun etc) it is evident that clusters consisting of two or more blocks tend to be heterogeneous in the sense

[8] The implication that elements taking the same slot are never used simultaneously is true. The inverse implication that elements taking different slots are always used simultaneously is false.

[9] It is plausible that the same mechanism — Barrier Rules — also triggers configurations with the so called late clitic placement, where the whole cluster attaches not to the first sentence category X but to some category Y to the right of X: [X]… Y = *a°* (1) =*b°* (2)=*c°* (3). Cf. [Zimmerling 2012] for discussion.

of (i). Indeed, since the main function of a Template Rule is to predict the order of any two clusterizing elements *x*, *y* and both of them may share the same taxonomic category C, the system needs additional parameters in order to generate the fixed order $x > y$[10]. In Cebuano and Bikol where clusters may consist of particles and pronouns the main principle of internal ordering is linked with the number-of-syllable-rule: monosyllabic elements precede disyllabic ones irrespective of their syntactic category. But if two monosyllabic particles or two monosyllabic pronouns form a cluster, the language still needs additional principles for their ordering. Therefore, a linearization algorithm for clusters must be set as a hierarchy of different principles.

	B. Prosodic Principle.	**C. Diachronic Principle:**
A. Categorial Principle. Cf. (vii) above.	B1. Number-of-syllable Rules: shorter monosyllabic elements precede longer disyllabic elements.	C1. Right-Edge Rules: new elements are added to the right edge of an already existing cluster.
	B2. Alternations of accented and deaccented elements in a cluster.	C2. Left-Edge Rules: new elements are added to the left edge of an already existing cluster.

Fig 3. Linearization Principles in clusters

The algorithms generating clusters may be analyzed in a diachronic aspect. The standard model discussed above capitalizes the situation where new clusterizing elements do not arise, so that the Template Rule at every moment of time signals that the sets of clusterizing and non-clusterizing elements do not intersect. It is necessary to analyze a broader class of situations, where new clusterizing elements may come to existence. There are three possible scenarios left.

A. Pure clitics join together and start clusterizing. This scenario implies that clusters always consist of clitics and obey HC. Initially, all clitics belong the set {- Cluster}, at some point t_0 some of them go over to the set {+ Cluster}.

B. Pure clitics attract other elements which start clusterizing. This scenario implies that existing clusters are expanded by new elements which before the moment t_0 did not belong to the set {+Clitic}. This may cause a violation of HC.

C. Non-clitic elements join together and become clusterizing. This scenario implies that some elements from the set {-Clitic} $\cup$ {- Cluster} at some moment

[10] We are not discussing situations when a word order system fails to generate the fixed order of two clusterizing elements and allows the variation $x > y \sim y > x$, since such situations are incompatible with the chosen definition of clusters. If one still wants to account for sporadic deviations from a Template Rule, it is possible to treat them as unstable transitional states of the system (linearization algorithm).

t_0 simultaneously go over to the set {+Cluster}. This may cause a violation of HC if some elements from the set {+Cluster} belong to the subset {-Clitic} and other elements from the set {+Cluster} belong to the subset {+Clitic}.

In a general form, Scenario A seems least likely. It is nevertheless taken as the default option in the Minimalist Program and related theories with the HC condition for clitics. Indeed, Scenario A is what one is inevitably left with if one adopts a HC-based synchronic analysis of clusters a la [Bošković 2001] or [Franks 2008]: both authors state that clusters consist of homogeneous clitics and though Franks identifies clitics as heads while Bošković treats them as left-branching elements, these framework-internal subtleties do not change the basic approach[11]. Scenarios B and C can be formalized if a cluster theory gives up the HC condition.

The choice between Scenarios A vs B and C may seem an abstract issue. I however argue that indirect evidence about the formation of clusters can be retrieved from the frequency rates of the clusterizing elements in a text corpus. I am proposing the following procedure: 1) Clauses with clusterizing elements are subdivided into two subcorpora — clauses with clusters (CLUSTER) and clauses with a single clusterizing element (SINGLE). CLUSTER and SINGLE share the same set of clusterizing elements <a,b,c…n>, where a,b, c … stand for slots in a Template Rule. 2) In clauses with clusters the frequency rates are measured for each clusterizing element. 3) Average frequency values are established for each subcorpus according to the formula m/n, where *m* is the total number of clusterizing elements in a subcorpus and *n* is the total number of Template slots that can be filled simultaneously in the same type of clauses. The m/n ratio predicts average frequency rates for each clusterizing element if all elements are represented evenly in a subcorpus. 4) The probability of clusterization is measured according to the formula $C/S = k_{CL}$, where C is the total number of clusterizing elements in CLUSTER, S is the total number of clusterizing elements in SINGLE and k_{CL} is the expected ratio of clusterization given that all clitics are represented evenly in CLUSTER. 5) The ratios $C_a/S_a = k_a$, $C_b/S_{b=} k_b$ …$C_n/S_n = k_n$ are calculated for every clusterizing element and the deviations from k_{CL} are measured. 6) Those clusterizing elements which show the positive difference $k_i > k_{CL}$ are identified as CORE if their frequency rates in both CLUSTER and SINGLE are significantly above the average values m_i/n_i. 7) Those clusterizing elements which have $k_j < k_{CL}$ are identified as ATTRACTED if their frequency rates in both CLUSTER and SINGLE are below the average values m_j/n_j and their absolute figures in CLUSTER and SINGLE are comparable. The CORE elements are more likely to have a higher frequency ratio in SINGLE, since they do not need clusterizing with other elements to be realized in a sentence. The ATTRACTED elements have a higher frequency ratio in CLUSTER, since they gain support from combinations either with CORE or other non-core clusterizing elements. The group ATTRACTED contains new elements added to an existing cluster; they tend to have a higher frequency in CLUSTER than in SINGLE, since their occurrences are supported by combinations with other elements. The proposed procedure proved

[11] With nearly the same effect one can analyze clusterizing clitics as pre-syntactic elements, the so called phrasal affixes which is made e. g. in [O'Connor 2002].

operational if statistics of clusters is used as an expert estimate confirming or testing linguistic hypotheses.

Conclusions

Clusters are formed according to Template Rules that arrange elements in an idiosyncratic order (ITP principle). Assumptions that all clusterizing elements are clitics and that clusterizing and non-clusterizing clitics represent the same set are not justified. Clusterizing elements belong to the same domain and level of representation (conditions ODP, CED and OLP). The Template structure can be analyzed in terms of slots and in terms of blocks of slots sharing the same taxonomic category. The TSP principle predicts that the number of slots in a Template equals the number of non-overlapping subsets A,B,C…N, so that each subset consists of elements a_1, a_2… a_n which stand in a complementary distribution with each other and are ordered uniformly respective any other elements from other subsets. Clusters meeting the ITP, ODP, CED, OLC and TSP may consist of heterogeneous elements. At the same time, adjacent clitic sequences which violate IPT, ODP, CED, OLC and TSP often meet the HC. The HC condition is incompatible with the analysis in terms of blocks, since the linearization algorithms which assemble clusters need a hierarchy of parameters, not a single parameter. Clusters typically consist of CORE and ATTRACTED elements: both groups have statistical profiles that can be retrieved by synchronic analysis. The CORE elements are older and/or functionally more prominent elements that have frequency rates above the average and a higher frequency in the subcorpus SINGLE. The ATTRACTED elements are newer elements that have frequency rates below the average and a higher frequency in the subcorpus CLUSTER. The proposed procedure of identifying CORE and ATTRACTED on the basis of their statistical profiles may complement linguistic analysis of clusters.

References

1. *Anderson S. R.* (1993), Wackernagel's Revenge: Clitics, Morphology, and the Syntax of Second Position. Language, Vol. 69, No.1, pp. 68–93.
2. *Bošković Ž.* (2001), On the Nature of the Syntax-Phonology Interface. Amsterdam: Elsevier.
3. *Ćavar D., Wilder C.* (1999). Clitic third in Croatian. Henk van Riemsdijk, ed. Clitics in the languages of Europe (Eurotype 20-5). Berlin- New York: Mouton, pp. 429–467.
4. *Franks S.* (2008), Clitic placement, prosody and the Bulgarian verbal complex. Journal of Slavic Linguistics 16/1, pp. 91–137.
5. *Guillaume A.* (2008), A Grammar of Cavineña. Berlin, New York: Mouton de Gruyter.
6. *Kosta P., Zimmerling A. V.* (2012), Slavic clitic systems in a typological perspective. Lilia Schürcks, Urtzi Etxeberria, Anastasia Giannakidou and Peter Kosta

(eds.) The Structure of NP and Beyond (Studies in Generative Grammar). Berlin, London: de Gruyter (to appear).
7. *Mushin I., Simpson J.* (2008), Free to bound to free? Interactions between pragmatics and syntax in the development of Australian pronominal systems. Language 84/1, pp. 566–596.
8. *Nash D. G.* (1986). Topics in Warlpiri Grammar. Cambridge, MA: MIT.
9. *O'Connor R.* (2002), Clitics and Phrasal Affixation in Construction Morphology. Proceedings of the LFG02 Conference. CSLI Publications, pp. 315–332.
10. *Zaliznjak A. A.* (1993), K izucheniju jazyka berestyanyh gramot [On the analysis of the language of birch bark letters]. Janin V. L., Zaliznjak A. A. Novgorodskie gramoty na bereste iz raskopok 1984–1989 [Novgorod birch bark letters from the excavations of 1984–1989]. Moscow: Nauka, pp. 191–319.
11. *Zariquiey Biondi, R.* (2011), A grammar of Kashibo-Kakataibo. La Trobe University, 2011.
12. *Zimmerling A. V.* (2002), Tipologicheskij sintaksis skandinavskih jazykov [Typological Scandinavian Syntax]. Moscow: Jazyki slavyanskoj kultury.
13. *Zimmerling A. V.* (2012), Sistemy porjadka slov s klitikami v tipologichsekom aspekte [Word Order systems with clitics in a typological perspective]. Voprosy jazykoznanija [Issues in linguistics], (in press).
14. *Zwicky A. M.* (1977), On Clitics. Bloomington: Indiana University Linguistics Club.

Abstracts

TUT 'HERE' AS A TEMPORAL PROXIMITY MARKER

Apresjan V. Yu. (valentina.apresjan@gmail.com), Higher School of Economics, Moscow, Russian Federation, Vinogradov Russian Language Institute, Moscow, Russian Federation

While space-time metaphor is a source of regular prepositional and adverbial polysemy in many languages, spatial deictic words develop temporal meanings much more rarely. However, space-time metaphor in deixis is not as exotic as it might seem, as demonstrated by the Russian spatial proximity marker tut 'here,' which develops a meaning of temporal proximity. Its meaning, though, is different from the meanings of classical deictic markers of temporal proximity, such as sejchas 'now.' Tut develops a synthetic meaning of actuality, which comprises the following semantic elements: (a) time period which includes the moment of speech, and such moments preceding and following it that are sufficiently close to the moment of speech to retain connection with it; (b) physical or mental space that includes the speaker; (c) a situation where the speaker is either a participant or an observer. Besides semantic differences, tut has a special semi-grammaticalized status (it is a particle turning into a semi-formed grammatical marker); communicative peculiarities (it cannot be part of a rheme or contrastivetheme); prosodic properties (it is a clitic).

RUSSIAN INTONATION CORPUS: A PRELIMINARY REPORT

Arkhipov A. V. (sarkipo@yandex.ru), **Zakharov L. M.** (leon@philol.msu.ru), **Krivnova O. F.** (okri@philol.msu.ru), **Kodzasov S. V.** (sankod@yandex.ru), **Lebedev A. A.** (andrettile@inbox.ru), M. V. Lomonosov Moscow State University, Moscow, Russia

The paper presents Russian Intonation Corpus (RINCO), the multimedia speech corpus designed for the study of prosody in Russian, its history, purpose, and functionality. RINCO started as a series of MS Access databases developed in 2004–2009 by a research team at the philological faculty of Moscow State University led by S. V. Kodzasov. The databases contained descriptions of most important prosodic characteristics (several tonal features, tempo, loudness, phonation types) for quasi-natural dialog utterances, as well as for a selection of spontaneous and prepared spoken narrative texts of different genres. These descriptions have been transferred to a multilayer time-aligned markup format, suitable for local use in ELAN annotator and for online browsing, playback and search using the LAT server platform. The corpus in its current state should be fully operational at http://languedoc.philol.msu.ru/rinco by the end of 2012. Various possibilities of enhancing the corpus are discussed, as well as some findings with a sample query for "IK-5" intonational construction.

TOWARDS A FREQUENCY DICTIONARY OF RUSSIAN IDIOMS

Baranov A. N. (baranov_anatoly@hotmail.con), **Dobrovolskii D. O.** (dm-dbrv@yandex.ru), **Kiseleva K. L.** (xenkis@mail.ru), **Kozerenko A. D.** (akozerenko@gmail.com), **Voznesenskaia M. M.** (voznes-masha@yandex.com), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russian Federation

Up to now we have not seen any frequency dictionary of Russian idioms. The reasons why this task seems so challenging are multiple. First of all, idioms appear in texts relatively rare if compared to "normal" words; next, their frequency depends heavily on the type of discourse as well as on the speaker's (or writer's) linguistic preferences. Our task is to compile a frequency dictionary of Russian idioms based on the list of idioms developed by our team for the Thesaurus of Modern Russian Idioms (2007). This list amounts to 8000 items and is being constantly revised and updated. We also have 4 text corpora elaborated especially for the needs of idiom analysis. They embrace the spheres of modern Russian relevant for idiomatic expressions: prose, drama, journalism, and detective fiction. We are going to take into account specifics of individual style of different authors: only those who use idioms regularly will be part of the text sample. The main type of information that the dictionary will provide is the frequency of each idiom in the corpus, taken as a whole, as well as in various genres. We will

also make a list of word forms that are part of the idioms and provide statistical information on the productivity of various semantic classes and grammatical forms in Russian phraseology. This parameter, quite novel for this field, was labeled phraseological activity. The results of this project can link together phraseology and other aspects of language system and facilitate contrastive idiom analysis.

PRELIMINARY CONSIDERATIONS TOWARDS DEVELOPING THE GENERAL INTERNET CORPUS OF RUSSIAN

Belikov V. I. (vibelikov@gmail.com), Russian Language Institute, RAS, **Selegey V. P.** (Vladimir_S@abbyy.com), ABBYY, RSUH, **Sharoff S. A.** (S.Sharoff@leeds.ac.uk), University of Leeds, RSUH

This talk presents the project for creating the General Internet Corpus of Russian (GICR). We start with analysing technological, structural, functional and content problems of existing Russian corpora. Then we discuss the need and the possibility for creating a new corpus which is based on the Russian Internet. Finally, we define the principles for text collection, classification and annotation, as well as the necessary parameters and functions of the interface of GICR.

TEXT SEGMENTATION IN THE OPENCORPORA PROJECT

Bocharov V. V. (bocharov@opencorpora.org), **Alexeeva S. V.** (alexeeva@opencorpora.org), **Granovsky D. V.** (granovsky@opencorpora.org), **Ostapuk N. A.** (nataxan90@gmail.com), **Stepanova M. E.** (mariarusia@gmail.com), **Surikov A. V.** (ksurent@opencorpora.org), OpenCorpora.org

By segmentation we mean text tokenization and sentence splitting. By default most text processing pipelines start from these two jobs before any morphology or syntax is invol ved. Moreover, later analysis stages depend on the way tokenization and sentence splitting are performed. In OpenCorpora project we have done manual tokenization and sentence splitting work on about 600K text forms.The paper presents a machine learning technique trained on this data, the results obtained, a segmentation standard and its motivation.

DESCRIPTION OF GAPPING IN A SYSTEM OF AUTOMATIC TRANSLATION

Bogdanov A. V. (bidon@inbox.ru), ABBYY, Moscow, Russia

In this paper we discuss a description of one type of ellipsis in a Russian-English and English-Russian automatic translation system. Short definition of this type of ellipsis (gapping) is given, as well as an overview about architecture of a system of automatic translation. The method of gapping description itself is considered and then its advantages and disadvantages are discussed in detail. Examples from real texts illustrating advantages as well as disadvantages of description are given.

A DICTIONARY OF DISCOURSE ELEMENTS OF RUSSIAN SPEECH: PROJECT DESCRIPTION (BASED ON CORPORA MATERIAL)

Bogdanova N. V. (nvbogdanova_2005@mail.ru), Saint-Petersburg State University

The paper is basically dedicated to the creation of Dictionary of Discourse Elements of Russian Speech. The author considers expanded hesitation constructions to be the markers of spontaneous speech. These constructions may consist of one or several words which actually fill pauses; we consider their occurrence to be discourse elements of a specific nature, typical for speech production. Our conclusions are based on observations of the Russian Speech Corpus (balanced annotated text collection and its component «One speech day»). The first glossary version of the dictionary and the structure of the dictionary entry are given.

AN AUTOMATIC MORPHOLOGICAL CLASSIFIER OF NOUN PHRASES IN RUSSIAN

Bolshakov I. A. (bolshakov34@mail.ru), Independent researcher, Moscow, Russia,
Bolshakova E. I. (eibolshakova@gmail.com), National Research University Higher School of Economics, Moscow, Russia

A morphological classifier of Russian noun phrases of almost arbitrary composition and length is described. Our study covers all peculiarities of Russian declination: fleeting vowels, multiplicity of declination patterns, animacy / inanimacy, specific additional cases, simultaneous occurrence of declinable nouns, adjectives and/or numerals in a phrase. The formula of Russian noun declination is a mere concatenation: word-form = pseudo-stem + pseudo-ending. The declination class is usually determined by the final 1 to 5 letters of the nominative. However exceptions are numerous and they are collected in many built-in lists. We consider separately isolated nouns, groups with no more than two declinable words, and extra long groups. The classifier is a part of CrossLexica, a large electronic dictionary of Russian, and is tested on 115,000 noun phrases of 1 to 6 words included in the CrossLexica dictionary. Based on the built-in classifier, an autonomous module is also constructed, which additionally declines extra-CrossLexica noun phrases.

DISCOURSE MARKERS USED FOR GOVERNING UNDERSTANDING OF TEXTS

Borisova E. G. (egbor@mail.ru), Moscow city teacher training university, Moscow, Russia

The article deals with speech particles (mostly modal particles and adverbs), that can be used by the Speaker for governing the Hearer in understanding speech. The units used for correcting or denying possible implications or presuppositions of the Hearer are investigated. These are particles like prosto (just), adverbs like sobstvenno (actually), voobsche (To it sum up) and some other. The investigation intends to proclaim the necessity of including discourse implications into the representation of the sense of an utterance or other discourse unit.

DATA-DRIVEN SPEECH PARAMETER GENERATION FOR RUSSIAN TEXT-TO-SPEECH SYSTEM

Chistikov P. G. (chistikov@speechpro.com), **Korolkov E. A.** (korolkov@speechpro.com), Speech Technology Center ltd, St. Petersburg, Russia

We propose a speech parameter generation approach for Russian based on hidden Markov models. The speech parameter sequence is generated from HMMs whose observation vectors contain speech characteristics. As a baseline we use the spectrum represented by mel-frequency cepstral coefficients (MFCC), pitch and duration parameters. All of them can be easily complemented by any other parameters, improving the quality. For the creation of the voice model we use linguistic and prosodic features which are the observations of every allophone in the utterance. This paper also presents the results of research into selecting the most effective features to characterize an allophone. Experimental results show that Russian speech can be successfully parameterized and an arbitrary utterance can be synthesized from the generated parameters.

RUSSIAN NATIONAL CORPUS AS A PLAYGROUND FOR SOCIOLINGUISTIC RESEARCH. EPISODE IV. GENDER AND THE LENGTH OF THE UTTERANCE

Daniel M. A. (misha.daniel@gmail.com), NRU HSE / Moscow State University, Moscow, Russia,
Zelenkov Yu. G. (yuryz@yandex.ru), Yandex, Moscow, Russia

The paper discusses sociolinguistic implementations of statistical analysis of the spoken subcorpus of the Russian National Corpus. Given the considerable size of the corpus (about 10 mln tokens), an analysis of co-variation of various linguistic parameters with one of the few sociolinguistic parameters available — the speaker's gender — may give rich and interesting results. One specific example of co-variation is considered in detail: the mean length of the utterance (in tokens). Comparing this parameter in public communication shows statistically significant difference between the speech of men and women (men talk more), while the same difference is absent in private communication. Another important parameter is the gender of the addressee. Again, co-variation is quite different in public and private discourse. In private communication, the utterances are longer when addressing someone of the same sex, the difference between men and women is not

statistically significant. In public communication, the utterances are longer when addressing a woman, whether the speaker herself is a man or woman. These conclusions are consistent with the results of sociolinguistic gender studies obtained elsewhere and by other methods. Linguistic difference between men and women are not absolute but depend on the communicative situation (public vs. private). Public discourse is a playground for linguistic competition in which men are the winning party. In private discourse, competition dissolves.

OPTIMAL FEATURE SELECTION FOR SPEAKER'S EMOTIONAL STATE CLASSIFICATION

Davydov A. G. (davydov-a@speetech.by), **Kiselev V. V.** (kiselev-v@speetech.by), **Kochetkov D. S.** (kochetkov-d@speetech.by), **Tkachenia A. V.** (tkachenia-a@speetech.by), LLC «Speech technologies», Minsk, Belarus

The research focus of present work is optimization of a feature set for voice emotion recognition. First part of the article contains a brief review of the most common speech features widely used in the emotion recognition tasks. It is shown that acoustic characteristics of a voice can be divided into five categories: prosodic, dynamic, qualitative, spectral and power. Also a number of the most effective features and statistical functionals derived from these features are discussed. After that two most widespread techniques of robust feature selection are explained. In the experimental part of the article, the feature selection algorithm developed on the basis of sequential feature selection (SFS) approach is presented. Further cross-validation procedure used in our studies is described. The recognition rate obtained on the Berlin database of emotional speech using optimized feature set (20 features) from the 10-fold cross-validation procedure is approximately 85%. In the conclusion we discuss some properties of the derived feature set and confusion matrix of the developed SVM classifier.

THE EFFECT OF THE SYNTACTIC ROLE OF THE ANTECEDENT ON AMBIGUOUS PRONOUN RESOLUTION IN RUSSIAN

Delikishkina E. A. (delik.kate@gmail.com), **Fedorova O. V.** (olga.fedorova@msu.ru), Lomonosov Moscow State University, Moscow, Russian Federation

According to a common view, referential choice is determined by the degree of activation of the referent in the cognitive system. It is even argued that our choice of referential device can be predicted solely on the basis of the activation factor. However, sometimes anaphora resolution is complicated by the presence of two or more competing referents, each of which enjoys activation high enough to be coreferent to the pronoun. If neither the context nor the grammar can help us identify the antecedent, we are confronted with referential ambiguity. It has long been the subject of debate what determines the choice of referent in such cases. Various factors have been shown to affect pronoun resolution. According to one account, the preferred antecedent of an ambiguous pronoun is the first-mentioned NP. The first-mention advantage is attributed to language-independent, general cognitive processes. Alternatively, it is argued by some researchers that pronoun resolution is affected by specific linguistic factors such as the antecedent's grammatical role. According to this account, the most favored antecedent is the grammatical subject. However, distinguishing between the two hypotheses using the data from languages such as English is impossible as the first mentioned referent is also usually a subject of the clause. In the present study we describe an experiment that has shown that speakers of Russian are more likely to corefer ambiguous pronouns to grammatical subjects regardless of whether they are mentioned first or not. We presented our subjects with ambiguous texts consisting of two sentences. The first one contained two referents, and the second one started with a pronoun that could refer to both of these referents. The first sentence had two versions: each of the referents could be either a subject or an object, which was attained by changing the diathesis of the transitive verb. Each of the 32 subjects that participated in the experiment was assigned to one of the two experimental lists that contained only one version of each stimulus text. The results showed that speakers of Russian considered the subject of the sentence to be an antecedent of the pronoun in 58,4 % of cases, which is statistically significant (binomial test, $p<0.01$), thus proving the importance and independence of the syntactic role factor. In conclusion, we reason that the apparent universality of the subject-preference strategy (the similar results were obtained for other languages, e. g. Finnish) supports the idea that the choice of grammatical structure is conditioned by the constraints on one's cognitive system, or rather working memory.

SYNONYMOUS FOCUS PARTICLES IN GERMAN AND RUSSIAN

Dobrovolskii D. O.(dm-dbrv@yandex.ru), **Levontina I. B.** (irina.levontina@mail.ru), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russian Federation

It is well-known that both Russian and German contain a considerable amount of particles of various types. In this paper, we discuss two German focus particles *eben* and *gerade*, as well as their Russian near-equivalents *imenno* and *kak raz*. Whereas *gerade* and *kak raz* can function as independent sentences, this way of use is blocked for *eben* and *imenno*. We show that this blocking goes back to some relevant semantic features of given particles. *Gerade* and *kak raz* mark the point of intersection of two situations, two mental schemas and the like. Although both *eben* and *imenno* may be used as separate utterances, this ability goes back to different semantic properties. The meaning of *imenno* is based on the idea of confirmation, of agreement with the interlocutor, while the German particle *eben* points to the most important, central element of the situation.

SUBJUNCTIVE IN COMPLEMENT CLAUSES

Dobrushina N. R. (nina.dobrushina@gmail.com), National Research University Higher School of Economics,Moscow, Russian Federation

Irrealis forms, such as subjunctive or conditional, are cross-linguistically widespread in complement clauses. Which verbs take irrealis complement clauses is subject to typological variation. I investigate which Russian verbs license subjunctive complementation. The list of predicates is compiled basing on the occurrences of the subjunctive forms in the Russian National Corpus. As a result, two types of subjunctive complement clauses are distinguished. For one group of verbs, subjunctive complementation is similar to purpose clauses. The verbs in the second, and much smaller, group, are used in constructions with low probability epistemic status.

FUNCTIONS OF FEEDBACK ITEMS *А*, *АГА* AND *ГМ* IN RUSSIAN PHONE CONVERSATION

Gerassimenko O. (olga.gerassimenko@gmail.com), University of Tartu, Tartu, Estonia

Functions of three feedback items in spontaneous speech are discussed on the basis of the preceding and following dialogical context. The analysis of 45-minute corpus of outpatients department phone calls shows that speakers use *гм* to signal the inapropriateness of the previous turn, *а* to signal the need for clarification and *ага* to signal the completeness and sufficiency of the previous turn.

AUTODEIXIS: TYPES AND MEANING

Grishina E. A. (rudi2007@yandex.ru), Institute of Russian Language RAS, Moscow, Russia

The study analyzes the main types of autodeictic gestures, i. e. the gestures which point to the speaker. The main configurations of the pointing hand are described and the main linguistic factors affecting these configurations are specified. We distinguish between the thematic autodeixis with the index finger and the performative autodeixis with the open hand. Besides, three types of autodeixes — portemanteaux have been described, which combine the deictic and non-deictic components. The study makes use of the data of the Multimodal Russian Corpus (MURCO).

A PROCEDURE FOR EVALUATING DEGREE OF CONVENTIONALITY OF METAPHOR EXPRESSIONS: FROM INTUITION TO OPERATIONAL CRITERIA

Gurin G. B. (grigory.gurin@uef.fi), **Belikova A. E.** (alex.belikova@gmail.com), University of Eastern Finland

The paper seeks to improve existing approaches to the quantitative analysis of metaphors in discourse and attempts to develop a new classification for metaphorically used lexical items based on the degree of their conventionality, i. e. on the extent of integrity of the original image element.

The novelty of the approach lies in its consistent avoidance of reference to intuitive and introspective criteria, which are of wide use in metaphor studies. The measurement of metaphorical creativity of the source and the target is based on a technique entailing a set of objectively established features, and referring mainly to external sources (dictionaries, corpora) and to a minor extent based on the intuition and subjective language competence of the researcher. The proposed classification goes in pair with a verifiable algorithmic procedure for identification of metaphor type. The method allows quantitative evaluation of the degree of metaphoricity for separate texts and text collections, as well as assessment of the metaphoricity of original texts and their translations. Russian is used as a source of illustrations.

BELARUSIAN AND RUSSIAN LINGUISTIC PROCESSING MODULES FOR THE SYSTEM NOOJ AS APPLIED TO TEXT-TO-SPEECH SYNTHESIS

Hetsevich Yu. S. (Yury.Hetsevich@gmail.com), **Hetsevich S. A.** (Novaeimya@gmail.com), **Lobanov B. M.** (Lobanov@newman.bas-net.by), United Institute of Informatics Problems of the National Academy of Sciences of Belarus (UIIP NAS Belarus), Minsk, Belarus

This paper describes the program NooJ, which provides a platform for the construction of modules for resolving linguistic problems in the area of text-to-speech synthesis. Belarusian and Russian are chosen as target languages. Basic and comprehensive electronic grammatical dictionaries for NooJ are described. We present the entire algorithm of converting dictionaries of the two languages into the form acceptable in NooJ, which retains all lexical, grammatical and accentual information. The dictionaries developed for NooJ help solve the problems of annotating words with lexical and grammatical categories and syllabic accents, as well as the problems of searching texts for a definite sequence of words according to their grammar and word forms. The grammars developed for NooJ are notable for their clarity and the speed at which they can be built. The structure and algorithm of the morphological grammar working on the text are given, including the localization of words written in Cyrillic and Roman letters. The structure and algorithm of the work of two syntactic grammars on a text are described: one aimed at searching for phonetic words and the other on searching for syntagms with a particular number of phonetic words. NooJ output produced after dictionaries or grammars have been applied to text are exported to text format Future work includes the completion of the transfer pf accentual information into Belarusian and Russian NooJ dictionaries, and construction of grammars for identifying accentual units in syntagms and grammars aimed at learning rhythmic structures of texts.

THESAURUS OF RUSSIAN EVERYDAY LIFE TERMINOLOGY: NEW PROBLEMS AND NEW TECHNIQUES

Iomdin B. L. (iomdin@ruslang.ru), **Lopukhina A. A.** (nastya-merk@yandex.ru), V. V. Vinogradov Russian Language Institute, Russian Academy of Sciences, **Piperski A. Ch.** (apiperski@gmail.com), M. V. Lomonosov Moscow State University, **Kiselyova M. F.** (mar-fed@yandex.ru), **Nosyrev G. V.** (grigorij-nosyrev@yandex.ru), **Rikityanskiy A. M.** (nordicdx@gmail.com), Yandex, **Vasilyev P. K.** (asthmor@gmail.com), **Kadykova A. G.** (annakadykova@nm.ru), **Matissen-Rozhkova V. I.** (heinin@mail.ru), Russian State University for the Humanities, Moscow

The paper addresses various issues associated with the development of a new encyclopedic thesaurus of Russian everyday life terminology (a current project by a group of researchers in V. V. Vinogradov Russian Language Institute, Russian Academy of Sciences). As new lexical material comes into the picture, lexicographers are faced with exciting challenges which require new approaches. Handling Internet data is one important issue. Frequencies of spelling variants and synonyms in texts of various genres, in particular, in blogs and user query logs, are determined using new Yandex-based tools specially designed for the project. The layout and the results of several kinds of speaker surveys and experiments are discussed in detail. The results of the paper may prove useful for a variety of lexicographic projects as well as for theoretical linguistics. The data under discussion, which has never been systematically studied before, appears to possess peculiarities worth further in-depth semantic research.

VERBS DESCRIBING SOUNDS OF INANIMATE OBJECTS: TOWARDS A TYPOLOGY

Kashkin E. V. (egorkashkin@rambler.ru), Moscow State University, **Reznikova T. I.** (tanja.reznikova@gmail.com), All-Russian Institute for Scientific and Technical Information, Russian Academy of Sciences, **Pavlova E. K.** (eliz-pavlova@yandex.ru), Moscow State University, **Luchina E. S.** (luchina-lena@yandex.ru), Moscow State University

The article deals, in a typological perspective, with verbs describing sounds of inanimate objects (cf. the noise of a door being opened, of coins in somebody's pocket, of a river, etc.). The analysis is based on the data from four languages (Russian, German, Komi-Zyrjan, Khanty), which were obtained from dictionaries, text corpora and field investigation. We discuss the primary meanings of these verbs and identify the parameters that underlie semantic distinctions between them (type of sound source and its features, type of situation causing the emission of a sound, acoustic properties of sounds). Then we concentrate on the semantic shifts undergone by sound verbs. First, we consider their metonymic changes, focusing on morphological and syntactic processes accompanying these shifts. Second, we analyze metaphoric uses of sound verbs, bringing out typical patterns of their derivation. These results should form the basis for a future large-scale typological investigation of sound verbs.

OPTIMIZING A MACHINE LEARNING BASED MODEL OF REFERENTIAL CHOICE

Kibrik A. A. (aakibrik@gmail.com), Institute of Linguistics, Russian Academy of Sciences, **Linnik A.** (skylinnik@gmail.com), **Dobrov G. B.** (wslc@rambler.ru), **Khudyakova M. V.** (mariya.kh@gmail.com), Lomonosov Moscow State University

In this paper we discuss different ways of how a machine learning based system of referential choice prediction may be optimized. Compared to the previous studies, we have improved and extended the annotation scheme. At the next step a "cheaper" set of parameters was applied in order to reach faster and less knowledge-rich processing. Our results demonstrate that only using the maximum of the available data the best accuracy of prediction can be gained, though it is possible to eliminate some of the highly "expensive" factors. Genre affiliation has been added to the system as one of the parameters and proved to increase the accuracy score. Finally, we have started a series of psycholinguistic experiments in order to explore the categorical vs. probabilistic character of the choice the speaker/writer makes. Our first results are promising in that in those instances in which the algorithms fail to make a precise prediction, more than one referential options are actually available, according to human judgments.

ADJECTIVES 'HEAVY' AND 'LIGHT' ON THE TYPOLOGICAL BACKGROUND

Kiuseva M. V. (pegas@nm.ru), **Ryzhova D. A.** (daria.ryzhova@mail.ru), **Kholkina L. S.** (kholkina@gmail.com), Lomonosov Moscow State University, Moscow, Russian Federation

The report is devoted to the study of Russian adjectives 'heavy' and 'light'. The unexpected symmetry of these lexemes is discussed: on the one side, they are antonymic practically in all meanings they have (internal symmetry), on the other side, this semantic area has the same structure in the languages that served the typological background for our research: Serbian, French, English and Chinese (external symmetry). Yet thorough research shows, that the similarity of lexemes has surface character. The following essential differences are revealed: 1. the adjective 'heavy' is used in direct meaning considerably more frequently than 'light'; 2. the adjective 'light' is used more frequently in metaphoric contexts and it particularly becomes apparent while expressing the meaning of degree: the meaning of down-toner of the lexeme 'light' is better developed than the meaning of intensifier of the lexeme 'heavy'; 3. adjective 'heavy' when it is used with certain nouns can involve the component of the meaning 'slow', while adjective 'light' can, on the one hand, involve the meaning with antonymical component 'fast', and on the other hand, through the meaning of down-toner, it can involve the component 'slow'; 4. analogical phenomenon with adjectives that have the meaning of 'light' can be seen in estimative component: in the whole the situation 'lightly' is rated positively, but there can be contexts in which the adjective with the meaning of 'light' has negative connotations. The adjective with meaning 'heavy' in Russian can only have negative connotations, but it can also develop positive connotations in other languages (e.g. 'important' in Chinese), if its original meaning is not 'difficult to lift or move', but '(objectively) weighing a lot'.

SOME DIFFERENCES BETWEEN CZECH AND RUSSIAN: A PARALLEL CORPUS STUDY

Klyueva N. M. (netsie@yandex.ru), Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic.

We present a comparative study of some constructions in Czech and Russian. Though Czech and Russian are closely related Slavic languages, they have a few differences at the level of syntax, morphology and their semantics. We discuss incongruities that we found in a parallel Czech-Russian corpus, mainly reflecting differences in the sentence structure. The linguistic evidence presented in the paper will be used while constructing the transfer module of a rule-based maachine translation system between Czech and Russian.

THE DATABASE "HUMAN BODY AND CORPOREALITY IN NATURAL LANGUAGE AND CULTURE (STRUCTURE, IDEOLOGY AND CONTENT)"

Klygina E. A. (katerinakk@rambler.ru), **Kreydlin G. E.** (gekr@iitp.ru), Russian State University for Humanities, Moscow, Russian Federation

A fundamental notion of "the semiotic conceptualization of the human body" has been proposed and investigated in order to describe the typical views of ordinary people on the human body, different phenomena of corporeality and interaction between verbal and nonverbal (corporeal) sign codes in an oral dialog. The authors present an original computer database holding information about diverse corporal (i. e. somatic) objects, their structural, physical and functional features and actions performed by the objects and on the objects. Internal structure of the database, its substantial online resources and interface that deal with different aspects of the Russian semiotic conceptualization of the human body and with the sign codes described are outlined and characterized. The database can answer different types of questions that are presented in the form of requests. The system has a user-friendly interface through which users can quickly and easily find answers to their questions and get results in a convenient form.

HUMAN CHARACTERS THROUGH THE PRISM OF ADVERBS

Kobozeva I. M. (kobozeva@list.ru), **Lukashevich N. Iu.** (natalukashevich@mail.ru), Lomonosov Moscow State University

The present paper deals with adverbs describing actions of a person from the character trait perspective — words like carelessly, tactfully, bravely, etc, derived from adjectives predicating to a person some character trait (careless, tactful, brave). The aim was to see what information such adverbs provide for defining patterns of behaviour that constitute the meaning of corresponding character trait adjectives. It is argued that the analysis of adverbs' contexts helps to outline more precisely the range of situations relevant for manifestation of a certain character trait, the range of actions which a person with the described character trait is inclined to perform, and their motivations. This is demonstrated by the analysis of contexts with sincerely, frankly and candidly from the British National Corpus.

CONSTRUCTING A LEXICO-TYPOLOGICAL DATABASE (FOR A STUDY OF PAIN PREDICATES)

Kostyrkin A. V. (languages@bk.ru), Institute of Oriental studies, RAS, **Panina A. S.** (panina-anna@yandex.ru), Institute of Oriental studies, RAS, **Reznikova T. I.** (tanja.reznikova@gmail.com), All-Russian Institute of Scientific and Technical Information, RAS, **Bonch-Osmolovskaia A. A.** (abonch@gmail.com), Higher School of Economics

We present a database developed for lexico-typological study of expressions of pain (demo version available at http://orientling.ru/bolit/). Its design implements the non-relational, NoSql approach, where data is organized into a flexible tree not limited in size and depth rather than presented as a table. Linguistic annotation is placed directly into the text of example sentences and their translations, so that in effect the database is structured as an annotated corpus. This formalism gives much freedom to both the developers in their task of annotating examples, and users in their queries, since it allows them to vary the level of detail according to how much information is available or needed. Linguistic annotation includes tags for syntactic roles, some syntactic con-

structions and their components (relative clauses, light verbs, formal subjects, parts of compound words), morphological information (tags for case, number, aspect etc), as well as semantic tags specific to the domain of pain (semantic roles and types of metaphoric shift).

THE RUSSIAN EMOTIONAL CORPUS: COMMUNICATION IN NATURAL EMOTIONAL SITUATIONS

Kotov A. (kotov@harpia.ru), National Research Centre "Kurchatov Institute", Moscow, Russia, **Budyanskaya E.** (davemonty2@gmail.com), Research Computing Center, M. V. Lomonosov Moscow State University, Moscow, Russia

The Russian Emotional Corpus (REC) includes annotated video recordingsof natural communication in tense emotional situations: oral university examsand talks between clients and officers at a municipal office regardingutility bills. Annotation of the corpus describes speech (text, syntagmaticstructure, speech acts, face threatening acts and the cases of irony), facialexpression, gaze direction and hand gestures. In the recorded emotionalsituations informants show diverse emotional cues and numerous rationaland emotional communicative strategies. The annotation allows us: (a)to search for a specific cue (like smiles or squints) and describe its usageand functions in natural communication, or (b) to search for specific patterns— like behavior markers of hesitation or facial cues at the end of utteranceswith face threatening acts. These observations help us to animateemotional computer agents, which receive semantic trees at their input,simulate rich emotional dynamics and produce semantic trees, ready-madephrases and gestures for the output.

REGIONAL PRONUNCIATION PREFERENCES FOR CERTAIN RUSSIAN AND FOREIGN WORDS

Krasnova E. V. (krasnova@speechpro.com), **Smirnova N. S.** (nsmirnova@speechpro.com), Speech Technology Center, St. Petersburg, Russian Federation

The aim of the reported study is to find out if any regional pronunciation preferences exist for certain words allowing of multiple pronunciations in Russian. In total 240 respondents from 12 Russian cities participated in the study. All of them were men aged from 20 to 60. Their pronunciations of 13 Russian words were recorded and analysed. Most of the pronunciation variants are at least tolerated by Russian orthoepic norm — in orthoepic dictionaries they are marked as either recommended or permissible or outdated. Auditory analysis was applied to the recorded speech data. It consisted in repetitive selective listening to speech segments of interest. Acoustics evidence was addressed in a limited number of controversial instances. As a result, pronunciation distribution statistics were obtained for the analysed words. On the whole, while designating some more or less clear tendencies, the results did not reveal any unambiguously regional pronunciation choices for the target words. Notably, some pronunciation options preferred by the overwhelming majority of speakers do not correspond to the recommended variant. Also of interest is the fact that there are speakers, although very few in number, who may pronounce the same word differently. The obtained statistics of frequency, regional distribution and within-speaker variability of analysed pronunciations can help in forensic comparisons of speech samples and can be taken into account in orthoepic dictionary compilation.

AUTOMATIC GENERATION OF EXTRACTION PATTERNS FOR SUBJECTIVE EXPRESSIONS FROM UNTAGGED TEXT

Kravchenko A. N. (etercian@gmail.com), Higher School of Economics, Moscow, Russian Federation

The goal of opinion mining is to extract and summarize opinionated contents from news, blogs, comments and reviews. One of the main tasks in opinion mining is detecting the boundaries of opinionated expressions and distinguishing between subjective expressions and factual information. High lexicon diversity for different domains excludes the possibility of formulating universal extraction rules that would work for any area of knowledge. In this paper we suggest a solution for this problem, reviewing a classification of subjective expressions in Russian and proposing an algorithm for automatic generation of extraction patterns for subjective expressions from untagged text based on label sequential rules (LSR). The algorithm also includes automatic tagging

of the training corpora and result filtering to minimize the need for human participation. At first the proposed algorithm uses an assortment of domain-independent pivots to distinguish opinionated sentences from the factual ones, which allows to avoid manual tagging. Possible subjective expressions are then extracted from selected sentences using a set of syntactic patterns. The applicability of this method is based on the fact that syntactic structure of subjective expressions is domain-independent as well. The resulting subjective expressions are, on the contrary, domain-specific. After that, the expressions are filtered with a use of probabilistic algorithm, increasing precision and therefore minimizing the need for human participation. The effectiveness of the proposed approach was evaluated on an collection of approximately 300 000 sentences, gathered from three different domains user reviews on movies, headphones and photo cameras. The best results (80% precision) were shown on domains with existing objective criteria and low lexical variability, such as reviews on cameras and headphones. For movie reviews precision reached 64,3% after filtering.

THE GENERAL CORPUS OF THE MODERN MONGOLIAN LANGUAGE AND ITS STRUCTURAL-PROBABILISTIC MODEL

Krylov S. A. (krylov-58@mail.ru), Institute of Oriental Studies, Russian Academy of Sciences, Moscow, Russia

The paper describes a General Corpus of the Modern Mongolian language (GCML), which contains 966 texts, 1 155 583 words. We also report a morphological analyzer for Modern Mongolian language (MML), a grammatical dictionary for 63 071 lexemes, a general table of morphological homonymy. The processor analyzes effectively 97 % of textual word forms which correspond to 76 % word forms from the inputs of the concordance to the GCML. MML can be described in its quantitative aspect, according to a structural-probabilistic model (SPM) of MML. SPM contains frequency dictionaries (FDs) of MML of different types: FDs of word forms, lexemes, grammatemes, root morphemes and allomorphemes, affixal morphemes and allomorphemes, flexionemes, grammemes. SPM allows describing behavior of various language units in the written text from the quantitative point of view: their frequency, distribution in texts, compatibility with other units etc. It is possible to transform the usual structural model into an SPM, which is based on statistical analysis of texts (in this model units of language are considered as possessing "the weight", the language oppositions and relations are being measured). The paper reports the top lists of some FDs: i. e. ranging FD of word forms (top-list of the upper 44 word forms having frequencies higher than 1700 ipm), ranging FD of lexemes (top-list of the upper 44 lexemes having frequencies higher than 2050 ipm) and ranging FD of grammatemes (top-list of the upper 44 grammatemes having frequencies higher than 2909 ipm).

WORDS WITH THE MEANING OF “MEDICINAL SUBSTANCE”: TODAY AND YESTERDAY

Krylova T. V. (ta-kr@yandex.ru), Institute of Russian language, Moscow, Russian Federation.

The paper deals with Russian words denoting medicinal substance: лекарство, препарат, средство. Our goal was to analyze the semantic relations in this sublexicon. We have found out that the sense of лекарство, the dominant and the oldest word of this group, is narrower than the sense of препаратand средство. (Particularly,лекарство denotes a medicinal substance, destined for the treatment of a disease, which is not obligatory for препарат orсредство).We conclude that препарат and средство are beginning to play the leading part in the semantic group of medicinal substances, which reflects the changes in the naive world model.

SEMANTIC TYPES AND SEMANTIC FUNCTIONS OF THE ADJECTIVIZED PARTICIPLES

Kustova G. I. (galinak03@gmail.com), Moscow State Pedagogical University, Moscow, Russian Federation

The paper explores correlations between an adjectivized participle and the corresponding verbal event, cf. *priglashennyj professor* (visiting professor) and the implications which qualitative meanings of the adjectivized participles are based on, cf. *povyshennyj interes* (heightened

interest), *produmannoe reshenie* (thought-out decision). Participials derived from the passive past participles are the most numerous class, whih includes a number of subclasses, such as:.1) Motivating event, cf. *razreshennyj miting* (permitted meeting). The event to which the participial refers does not affect the "visible" state of the object. It is impossible to reconstruct such an event, it is necessary to know about it. Thus, the motivating event gives a sign sufficient to identify a subclass of the general class. 2) Perfect, cf. *okrashennye volosy* (dyed hair). This is the traditionally understood perfect — the state as a result of a prior event. 3) Comparative model, cf. *povyshennye trebovanija* (increased demands). This is a large class of participials, consisting of two main subclasses. The first subclass includes participials from verbs, in their turn formed from adjectives (sometimes from comparatives like *luchshe* 'better', *men'she* 'less'): *zamedlennoje razvitie* (delayed development) — 'slower than normal', *zatemnennye stekla* (tinted glasses). Participials of the second subclass denote a modified shape: *zakruglennyj konchik* (rounded tip), *iskrivlennyj palec* (twisted finger). Such participials express the idea of a relatively small shift in the attribute or parameter scale. 4) Quantitative model ("sufficient degree"), cf. *nasyshchennyj rastvor* (saturated solution). This group includes participials with a component 'high degree'. This component turns a participle into a qualitative adjective. 5) Qualitative model ("weak evaluation"), cf. *prisposoblennoe pomeshchenie* (adapted premises). Participials of this group have a component 'good' or 'bad', so that *prisposoblennoe pomeshchenie* implies well adapted premises, *zapreshchennyj priem* (forbidden method) implies a foul, bad method'. Adjectives *prekrasnyj* (beautiful), *uzhasnyj* (terrible) etc. clearly belongs to the class of estimates, their lexical meaning is reduced to the evaluation. Hpwever, participials like *prisposoblennyj* (adapted) or *zapreshchennyj* (forbidden) should best be classified as positive / negative attributes than real estimates. Although participials often have semantic counterparts among ordinary verb adjectives they contain important interpretation models not found in other classes. Participials fill in the gaps of characteristics of objects.

RUSSIAN LEARNER PARALLEL CORPUS AS A TOOL FOR TRANSLATION STUDIES

Kutuzov A. B. (akutuzov72@gmail.com),Lionbridge Technologies, Inc, Tampere, Finland,
Kunilovskaya M. A. (mkunilovskaya@gmail.com), **Oschepkov A. Y.** (aoschepkov@gmail.com),
Chepurkova A. Y. (AnnaChepurkova@yandex.ru), Tyumen State University, Tyumen, Russia

The paper presents a project aimed at the development of a Russian Learner Parallel Corpus, discusses the existing analogues, describes the current status and the tasks in which it could be used. The existing parallel corpora contain (comparatively) "correct" translations; whereas the aim of the present project is to create a sufficiently large corpus of imperfectly translated Russian and English texts together with their sources and use it as a tool for translation studies, especially those related to translation mistakes. The new corpus will be a valuable resource for computational linguistics as it provides another way of getting data for evaluation which could be used to improve machine translation systems. As of now, the corpus is available online, it already contains nearly half a million word tokens and is growing. The main source of material is translations made by student translators in Russian universities.

BUILDING THE INVENTORY OF RUSSIAN NOMINAL CONSTRUCTIONS

Lashevskaya O. N. (olesar@gmail.com), National Research Institute Higher School of Economics, Moscow, Russia, **Mitrofanova O. A.** (alkonost-om@yandex.ru), **Grachkova M. A.** (maaag86@mail.ru), Saint-Petersburg State University, Saint-Petersburg, Russia, **Romanov S. V.** (complefor@rambler.ru), ZAO «Internet-Projects», Saint-Petersburg, Russia, **Shimorina A. S.** (shinas@yandex.ru), Institute of Linguistic Studies RAS, Saint-Petersburg, Russia, **Shurygina A. S.** (sanyana@gmail.com), Russian State University for Humanities, Moscow, Russia

The paper presents experimental results of automatic construction identification performed on the Russian National Corpus (RNC). For this purpose we developed a toolbox which allows extraction and processing of co-occurrence data from RNC samples. Russian nouns are chosen as target words. Lists of constructions were built for each target word. By constructions we mean frequent word combinations which include a target word and frequent lexical-semantic tags — context marker of certain meanings of a target word, as well as frequent lemmas representing the given lexical-semantic tags. E.g.: ВИД (kind, sort, type) + r:abstr t:sport: спорт (sport), футбол

(football), биатлон (biathlon), etc. Extracted constructions are grouped according to their structure and lexical-semantic content. In conclusion we verify the experimental results. which implies comparison of lists of constructions with lists of collocations, idioms, etc. registered in various linguistic resources (bigram search engines, dictionaries).

THE IMPACT OF SPEECH AND SPEAKER CHARACTERISTICS ON THE ACCURACY OF AUTOMATIC SPEECH RECOGNITION

Liudovyk T. V. (tetyana.lyudovyk@gmail.com), International Research/Training Center for Information Technologies and Systems, Kyiv, Ukraine

The paper reports an investigation of effects caused by speech style and speaker characteristics on speech recognition accuracy. Results of Ukrainian read and spontaneous speech recognition are analyzed. The speech material consists of broadcast news (15 %), talk shows (29 %), and real court reports (31 %). The test corpus counts 17 000 words and includes speech of 9 male and 9 female speakers. The speakers are TV news presenters, politicians, journalists, lawyers, and "ordinary" members of trials. Two experiments have been conducted. The first one consists in automatic speech recognition, while the second one is based on annotations of speech made by experts. Different speech characteristics are investigated, namely speaking rate, speech disfluences (breathing, hesitation fillers, restarts, fragmented words, reduction and prolongation of words), slang and colloquial words. The most error-prone speech characteristics are restarts and fragmented words, slang and colloquial words. Breathing and hesitation fillers are less error-prone because they are successfully modeled in the speech recognition system. Fast read speech is recognized more accurately than slower spontaneous speech of the same speaker. The more accurate recognition (word error rate 10–15 %) is achieved for speakers who have experience in oral presentations. The obtained results allow roughly prediction of the accuracy of speech recognition based on such speaker experience. Ukrainian speech recognition accuracy achieved so far allows using the speech recognition system for automatic transcription (subtitling) of broadcast news.

ORAL SPEECH AND INSTANT MESSAGING: THE IMPACT OF THE COMMUNICATION CHANNEL ON DIALOGUE STRUCTURE

Liusina V. S. (krotcaja@gmail.com), Institute of Linguistics of the Russian Academy of Sciences, Moscow, Russian Federation

Correlation between oral and virtual dialogues is analyzed. It is known that dialogues consist of minimal dialogic units (MDU). Contrary to a wide-spread opinion that an MDU cannot include another MDU, there are evidences that this is possible when one of the dialogues is not intentional. The structure of 60 Russian Internet dialogues is analyzed. If these dialogues satisfy the "close-to-oral-conversation" conditions (i.e. the interlocutors are not involved in any other mental activities while talking), a mutation of dialogue structure occurs. There are five steps of mutation that lead from an overlooking of a remark to a communicative failure. In such "mutated" dialogues, two interlaced "intentional" MDU are present. However, the communicative failure occurs very rarely due to multitasking. Normally, people are involved in several activities while talking via Internet, which provides necessary pauses in Internet conversation. It can be concluded, therefore, that Internet conversation is a wholly new type of communication.

ON THE STRUCTURE OF A THESAURUS OF PROVERBS

Malkova A. S. (asmalkova@gmail.com), Russian State University for the Humanities, Moscow, Russian Federation

The semantics of a proverb corresponds to some class of situations in which this proverb can be used. We argue that the investigation of the structure of proverb situations can give some productive ideas of how to set up the proverb thesaurus (i.e. a collection of proverbs where each item is supplied with multiple cross-references to relevant texts). We suggest using a special meta-language in order to represent proverb situations. The main items of this meta-language are concepts from the domains of a) human acts and characteristics like labor, help, knowledge etc. and b) outer circumstances caused by human acts and characteristics like success, profit, pleasure etc. Each

concept has a dual structure and forms an opposition with a positive and a negative part (such as labor-idleness, help-harm, knowledge-ignorance; success-failure, profit-loss, pleasure-suffering etc.). Two oppositions set together make an elementary statement, where a) one opposition is seen as a result of the other or b) the priority of one opposition over the other is set up. To specify the priority, the following pairs can be used: unreal-real, wanted-accessible, temporary-constant, mundane-hollow etc. If each proverb in the collection is provided with such a formal description, cross-references are created automatically, since semantic similarities will be an evidence of coincidences in the formal structure. The suggested procedures were verified on an experimental proverb corpus, which was not so big (500 items) but included most of generally used texts. Project web-site: http://metaphora2.ru/.

DEVELOPMENT OF CHINESE LANGUAGE LEXICAL-SEMANTIC DICTIONARY FOR THE MULTI-LANGUAGE NLP SYSTEM

Manicheva E. S. (Ekaterina_M@abbyy.com), **Dreyzis Yu. A.** (Yulya_D@abbyy.com), **Selegey V. P.** (Vladimir_S@abbyy.com), ABBYY, Moscow, Russia

This paper deals with bilingual lexical-semantic dictionary of Mandarin Chinese designed for NLP purposes. This dictionary of Chinese core vocabulary has been compiled according to the principles of model-based universal multi-language linguistic technology Compreno, developed in ABBYY. Nowadays most of lexical data-bases of Mandarin Chinese are based on WordNet principles. Our work shows that Chinese language might as well successfully fit in an alternative universal lexico-semantic database. Here we present an overview of major methodological challenges and solutions to integrate Chinese language data into Compreno framework. At the moment lexical-semantic dictionary of Mandarin Chinese covers more than 8000 meanings with well-structured comprehensive information on deep semantic and syntactic model of a meaning, its lexical and grammatical co-occurrence restrictions, and further work on dictionary is still going on. This paper focuses on typological differences between Chinese and European languages in terms of basic unit for dictionary entry, grammar paradigm of a word and its meanings, differences in syntactic realizations of deep semantic model. The paper also gives reasons why certain theoretical approaches prevailing in Chinese linguistic tradition were revised to serve better application needs, what principles of meaning definition were taken into consideration to provide detailed and complete lexicographic descriptions, compare and contrast Chinese with Russian or any other language.

SHOLOKHOV, OR KRYUKOV AFTER ALL? NONFORMAL PROCEDURES TO IDENTIFY THE AUTHORSHIP OF “TIKHIJ DON”

Mikheev M. Yu. (m-miheev@rambler.ru), NIVC MGU, Moscow

The article deals with an authorship identification issue. A variety of texts by Sholokhov and Kryukov are analyzed by context comparison. For each hypothetical coincidence of key words a formal sample is chosen. Only context structures of Sholokhov and Kryukov correspond to these samples. The search of coincidences in the Russian National Corpus and in the full texts using Google helps verify the unique character of variation. Then we examine some keywords, isolated expressions and whole fragments from K. and Sh. texts, which represent the closest duplicates of one other. As a result, we get different types of coincidences, single and punctual or multiple (consisting of a set of structures), but also those distributed among various places in the texts of both authors. In each case we evaluate the probability that the original author of the expression could be Kryukov.

ESTONIAN SPEECH SYNTHESIS: APPLICATIONS AND CHALLENGES

Mihkla M. (meelis@eki.ee), **Hein I.** (kiisu@eki.ee), **Kalvik M.** (mariliis@eki.ee), **Kiissel I.** (indrek@eki.ee), **Tamuri K.** (kairi.tamuri@eki.ee), Institute of the Estonian Language, Tallinn, Estonia, **Sirts R.** (risto.sirts@err.ee), Estonian Public Broadcasting, Tallinn, Estonia

In the 21st century Estonian speech synthesis has been developed using the more widespread methods and freeware development systems (MBROLA, Festival, eSpeak, HTS). The applications have hitherto been developed mainly in view of the needs of the visually impaired (audio system for reading electronic texts, voicing of subtitles, creation of audiobooks). The major challenges currently facing the Estonian specialists are naturalness of the output speech and expressive

speech synthesis. The article is concerned with the issues of statistical modelling of the prosody of synthesized speech and the relations of prosody with other language levels as well as with extra-linguistic features. Analysis of the emotion-bound acoustic parameters (pauses, speech rate, formants, intensity and pitch) enable one to model emotions for speech synthesis. In addition, speech synthesis interfaces are discussed. By means of such interfaces users could control the process of speech synthesis, monitor text-to speech transformation, follow text structure and vary the parameters (voice loudness, speech rate, voice pitch) of the synthetic voice in various voice applications.

WHAT DO PEOPLE ASK THE COMMUNITY QUESTION ANSWERING SERVICES AND HOW DO THEY DO IT IN RUSSIAN?

Mukhin M. (mfly@sky.ru), Ural Federal University, Yekaterinburg, Russia, **Braslavski P.** (pbras@yandex.ru), Kontur Labs/Ural Federal University, Yekaterinburg, Russia

In our study we surveyed different approaches to the study of questions in traditional linguistics, question answering (QA), and, recently, in community question answering (CQA). We adapted a functional-semantic classification scheme for CQA data and manually labeled 2,000 questions in Russian originating from Otvety@Mail.Ru CQA service. About half of them are purely conversational and do not aim at obtaining actual information. In the subset of meaningful questions the major classes are requests for recommendations, or *how*-questions, and fact-seeking questions. The data demonstrate a variety of interrogative sentences as well as a host of formally non-interrogative expressions with the meaning of questions and requests. The observations can be of interest both for linguistics and for practical applications.

APPLICATION OF N-GRAMS AND OTHER LETTER- AND WORD-LEVEL STATISTICS TO SEMANTIC CLASSIFICATION OF UNKNOWN PROPER NOUNS

Nekhay I. V. (nekhayiv@gmail.com), Department of image recognition and text processing, DIHT MIPT, Moscow, Russian Federation

Automatic semantic classification of unknown proper nouns is a significant problem in the field of automatic text analysis. We investigate the ability to classify proper noun phrases frequent in Russian language using machine learning approaches and n-gram statistics, as well as other letter- and word-level feature statistics such as word capitalization, number of words, presence of abbreviations and numbers, etc. We use only internal information, but, as we expect to use external information in future, we test scenarios that can arise when the former is utilized for classification. We show that such simple features allow to achieve accuracies up to 80-99,9% in 1-1 trials, 89-99% in 1-other trials and 85% accuracy in 5-way trial in 5 pre-selected categories, using training set of about 2000 examples in each category. These results come very close to results of (Patel & Smarr, 2001), achieved by a similar system designed to classify English proper nouns.

COMBINING MULTIPLE FEATURES FOR SINGLE-WORD TERM EXTRACTION

Nokel M. A. (mnokel@gmail.com), **Bolshakova E. I.** (eibolshakova@gmail.com), Moscow State University, Moscow, Russia, **Loukachevitch N. V.** (louk_nat@mail.ru), Research Computing Center, Moscow State University, Moscow, Russia

The paper describes experiments on automatic single-word term extraction based on combining various features of words, mainly linguistic and statistical, by machine learning methods. Since single-word terms are much more difficult to recognize than multi-word terms, a broad range of word features was taken into account, among them are widely-known measures (such as TF-IDF), some novel features, as well as proposed modifications of features usually applied for multi-word term extraction. A large target collection of Russian texts in the domain of banking was taken for experiments. Average Precision was chosen to evaluate the results of term extraction, along with the manually created thesaurus of terminology on banking activity that was used to approve extracted terms. The experiments showed that the use of multiple features significantly improves the results of automatic extraction of domain-specific terms. It was proved that logistic regression is the best machine learning method for single-word term extraction; the subset of word features significant for term extraction was also revealed.

BASHKIR INTERNET: THE VOCABULARY AND PRAGMATICS IN THE QUANTITATIVE ASPECT

Orekhov B. V. (nevmenandr@gmail.com), Russian Presidential Academy of National Economy and Public Administration, Bashkir State University, **Gallyamov A. A.** (azamat.gallyamov@gmail.com), Bashkir State University

The paper deals with a quantitative aspect of the Bashkir language segment of the Internet. We analyze the results of a special crawler's work. Our crawler has indexed Bashkir sites and collected the linguistically valuable data on word form frequencies. This data differs from word frequencies in Bashkir printed texts or in the Russian Internet. Most of the frequent words are marked as official. Obscene words are almost nonexistent. This means that the Bashkir Internet is not designed for any kind of communicative needs; the main goal seems to convey the message of the existence of the Bashkir language and its presence on the Web. Internet terms like "site" and others are rare in the Bashkir web segment. Such popular words in Runet as "job", "mobile phone" and others are not as frequent in the Bashnet.

METHODS OF LEXICOGRAPHIC DESCRIPTION OF RUSSIAN SPONTANEOUS SPEECH (A CORPORA RESEARCH)

Os'mak N. A. (natalia.osmak@yandex.ru), Saint Petersburg State University

The paper presents the results of an analysis of possible methods of lexicographical description of Russian spontaneous speech aimed at the creation of a General Russian Conversational Dictionary. The analysis is based on the material of the Corpus of Spoken Russian "One Speech Day", which presents unique linguistic material, enabling fundamental research in many aspects: study of real spontaneous speech, phonetics and grammar of spoken language, psycholinguistics, communication studies, etc. The General Russian Conversational Dictionary can be created through the description of most frequent units or through the analysis of lexical semantic groups. Advantages and disadvantages of the two methods are analyzed. Tentative lexicographical description of a lexical semantic group representing nominations of humans are given.

COMMUNICATIVE PERSPECTIVE INTERPRETATION: BASIC STRUCTURES AND LINEAR-ACCENTUAL TRANSFORMATIONS

Paducheva E. V. (elena.paducheva@yandex.ru), VINITI, Russian Academy of Sciences, Moscow, Russia

Communicative structure (C-structure), which is also called communicative perspective of a sentence, relies upon linear-accentual structure (LA-structure), which is, basically, a linear sequence of tonal groups, each with its syntactic and, possibly, lexical characteristics. C-structure identifies each group as a constituent of a certain kind, such as theme or rheme, and posits semantic relationships between these constituents. These relationships can be hierarchical; for example, a constituent theme provides a place for theme-rheme division of the lower level. The problem is to create a calculus of possible C-structures of Russian sentences and provide each C-structure with the description of its contribution to communicative semantics of the corresponding sentence. A kind of transformational approach to the problem is accepted. Several basic (or neutral) C-structures are revealed, which most naturally correspond to lexico-syntactic structure of the sentence, and a set of LA-transformations is presented. For each transformation its contribution to communicative meaning of the sentence is described.

PROBLEMS AND METHODS IN ANALYSIS OF RUSSIAN TEXTS IN PRE-REFORM SPELLING

Polyakov A. E. (pollex@mail.ru), Ushinsky State Scientific Pedagogical Library, Moscow, Russia

The existing linguistic processors (spellcheckers, lemmatizers, OCR programs) are not suitable for analysis of pre-reform Russian texts because of numerous graphical, morphological and lexical differences from the modern Russian language. Some lemmatizers have a restricted support of pre-reform spelling, but they are closed source and cannot be modified or extended. We have de-

veloped a lemmatizer which can properly analyze pre-reform texts and has a facility for flexible adaptation to other spelling systems. This paper discusses the problems in the analysis of pre-reform Russian texts (obsolete forms, spelling variants) and the methods of their solution (normalization, modification of the grammatical model, etc.).

VARIATION IN GENDER FORMS IN THE GROUP OF PLURALIA TANTUM IN THE RUSSIAN LANGUAGE

Savchuk S. O. (savsvetlana@mail.ru), Vinogradov Institute for the Russian Language, RAS

The paper presents the results of a corpus-based study on variation in Russian *pluralia tantum* nouns. Most of them vary in gender and/or genitive plural forms. The list of variants was composed by analyzing textbooks and dictionaries compiled at the beginning and the second half of the 20th century. The variants are classified according to their morphological and semantic features. The usage of every lexeme from the list was analyzed in the texts of the Russian National Corpus, all variants were registered in the database and the correlation between variants was determined. The comparison of corpus data with the data derived from dictionaries made it possible to detect the changes in correlation between variants within the studied period and to formulate some trends in variants functioning. The results of the research are used to correct morphologic annotation; they can also be regarded as a material for the lexicographic description of variants.

TOWARDS THE AUTOMATED ENRICHMENT OF MULTILINGUAL TERMINOLOGY DATABASES WITH KNOWLEDGE-RICH CONTEXTS

Schumann A. K. (anne.schumann@tilde.lv), Tilde SIA/University of Vienna, Riga/Vienna, Latvia/Austria

This paper describes ongoing Phd thesis work dealing with the extraction of knowledge-rich contexts (KRCs) from specialized Russian and German text corpora for the semantic enrichment of terminological resources. In recent years, automatically extracted KRCs have been proposed as a means for deriving empirically grounded concept descriptions for terminography while maintaining the time and costs spent for the acquisition of such descriptions on a reasonable level. KRCs have been studied for a number of European languages ranging from English over French and Spanish to Catalan, however, not much effort has yet been put into widely spoken, but typologically different languages such as Russian or German. This paper, therefore, describes research efforts aiming at the extraction of KRCs in Russian and German for the purpose of termbase enrichment. Section 1 of this paper presents a brief introduction to KRC research and the motivation for this study. Section 2 gives an overview over related work. Section 3 describes the KnowPipe KRC extraction framework, whereas section 4 outlines ranking experiments with KnowPipe on Russian and German data. Section 5 summarizes the results and describes future work.

ON THE METAPHORA AND THE METONYMY OF THE RUSSIAN PARAMETRIC NOUN

Semenova S. Iu. (Sonya_sem@mail.ru), Institute of Scientific Information on Social Sciences, RAS, Russian State University for Humanities, Moscow, Russian Federation

The paper is concerned with ambiguity and semantic transfers of the Russian parametric noun. The semantic processes are explicated through changes in the argument taxonomy of this predicate noun. We consider noun phrases with the genitive case of the parametric noun argument Metaphoric transfers of the parametric noun associated with disappearance or alteration of the quantitative meaning and measurability are essential for NLP tasks, namely, for automatic extraction of the parametric data. The concepts of weak and strong metaphor are introduced. The weak metaphor that maintains the measurability seems to be one of the principal ways of terminology formation. Metonymic shifts of the parametric noun and its argument should be fixed and ranked in the NLP-oriented dictionary.

RESEARCH OF PLAGIARISM IN THE STUDENT WORKS

Sharapov R. V. (info@vanta.ru), **Sharapova E. V.** (mivlgu@mail.ru), Murom Institute of Vladimir State University, Murom, Russian Federation

In this paper we study plagiarism in student works. We consider as plagiarism a direct unattributed copy of a text. We describe how students try to hide plagiarism. We have analyzed a collection of tests, course papers, and student working papers. When writing their own work (a course paper or a student working paper), students usually copy material from just one source, while when writing a test, they usually use more sources. We found that the methods used to hide the plagiarism usually include changing of text; reduction of text; changing of certain pronouns, adjectives and verbs; replacement of Russian letters by similar Latin ones. Among the less frequently used methods are replacements of parts of text, changed punctuation, substitution of invisible characters for spaces, manual and automatic substitution of synonyms, copying texts from some sources and changing them. Methods of hiding plagiarism may change over time if they become less effective.

CORPUS OR EXPERIMENT?

Shmelev A. D. (shmelev.alexei@gmail.com), Moscow Pedagogical State University, Moscow, Russia; Vinogradov Institute of Russian Language, Russian Academy of Sciences, Moscow, Russia

The paper discusses general distinction between observation and experiment as applied to linguistic research. It claims that the resolution of certain issues has to be based on observation, that is, on corpus data (e. g., studies of ancient languages) while other issues require experiment (in particular, issues concerning the difference between common usage and linguistic standards). Two main varieties of experiment may be distinguished, namely, experiment on the researcher's linguistic competence and interrogation of informants. The latter is the only accessible experimental method if the researcher has no linguistic competence of the language or dialect under investigation. To illustrate the point, the paper uses the example of regional linguistic standards (in particular, norms of pronunciation). It discusses various pitfalls on the way to an accurate account of regional linguistic standards. To avoid falling into one of those pitfalls, the researcher should be clear in his/her mind about the ultimate objective of the investigation and word questions to the informants in a clear form if s/he is going to use a questionnaire.

RUSSIAN-ENGLISH THESAURUS ON COMPUTATIONAL LINGUISTICS

Sokolova E. G. (minegot@rambler.ru), Russian State University for the Humanities, Moscow, Russia, **Kononenko I. S.** (irina_k@cn.ru), **Ershov A. P.** Institute of Informatics Systems SB RAS, Novosibirsk, Russia

This paper summarizes the experience in the construction of Russian-English information retrieval thesaurus on Computational Linguistics (CL). The need for relating thesaurus terms to the subareas of CL and adjacent sciences is substantiated and the hierarchical structure of subareas is discussed. The kinds of information given in the thesaurus term entry are outlined. A number of terminology description issues are discussed with regard to the specific features of the constructed thesaurus such as bilinguality and insufficient development of Russian CL. Terminological problems are analysed using classification parameters.

AUTOMATIC GENERATION OF TEXT CORPORA FOR CREATING VOICE DATABASES IN A RUSSIAN TEXT-TO-SPEECH SYSTEM

Solomennik A. I. (solomennik-a@speechpro.com), Speech Technology Ltd., Minsk, Belarus, **Chistikov P. G.** (chistikov@speechpro.com), Speech Technology Center Ltd, St .Petersburg, Russia

This paper deals with the problem of speech database design for the needs of unit selection text-to-speech synthesis. An obligatory condition for the naturalness and intelligibility of synthesized speech is a high quality speech database. We propose a computer program developed specifical-

ly for the Russian language which creates a phonetically balanced text corpus of given size. We present a description of the program and a comparison of an automatically constructed corpus and some arbitrary corpora. The automatic text corpus generation program is part of a new voice building system for VitalVoice Russian TTS. It helps to supplement a text corpus with missing phonetic units. Further possible improvements of the algorithm are also discussed. We consider several ways to take into account intonational variation of units in a database at the stage of the preparation of a text corpus.

USING SENTIMENT-ANALYSIS FOR TEXT INFORMATION EXTRACTION

Solov'ev A. N. (a.solovyev@i-teco.ru), **Antonova A. Ju.** (a.antonova@i-teco.ru), **Pazel'skaia A. G.** (Pazelskaya@i-teco.ru), «I-Teco», Moscow, Russia

This article aims to demonstrate relationship between sentiment-analysis and information extraction from text. To illustrate the idea we propose a sentiment-based summarization prototype instrument. The performance of the method in comparison with 3 another summarization methods is tested on a small corpus of mass media and blog posts.

CONJUNCTIONS, CONNECTORS, AND VALENCY THEORY

Uryson E. V. (uryson@gmail.com), V. V. Vinogradov Russian Language Institute (RAS, Moscow, Russia)

According to traditional grammar conjunctions are words connecting clauses (or words / word combinations). Still many conjunctions can also connect sentences and longer fragments of texts, being in fact connectors, (these are "soft" conjunctions), while some conjunctions are able to connect only clauses (these are "hard" conjunctions). Many Russian conjunctions, both coordinating and subordinating, are soft, but there are also hard coordinating and subordinating conjunctions. Hard conjunctions are easily described in the frameworks of valency theory as predicates with (at least) two semantic and syntactic actants. Soft conjunctions and connectors also have two semantic actants. But at least one of these actants cannot be revealed in a text by an algorithm. It is argued that some refinements of valency theory are necessary for representing syntactic properties of connectors and soft conjunctions.

GOVERNMENT OF THE BORROWED NOUNS DENOTING PROFESSIONS IN THE MODERN RUSSIAN LANGUAGE

Valiakhmetova A. R. (episteltjes@gmail.com), V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russian Federation

Linguists who study borrowed words concentrate either on sociolinguistic problems or the process of the foreign words adaptation. As for their syntactic features, they are usually left unconsidered. The present paper deals with the semantics and government of the borrowed nouns denoting professions in the modern Russian language, namely: 'manager', 'agent', 'freelancer', 'outsourcer', 'designer', 'couturier', 'stylist', 'coach' and 'tutor'. These nouns can be divided into three semantic groups: 'employees' ('manager', 'agent', 'freelancer' and 'outsourcer'), 'artists' ('designer', 'couturier' and 'stylist') and 'teachers' ('coach' and 'tutor'). There are two government variants of these words: non-prepositional genitive and prepositional phrase *po* + dative. Different factors affecting the usage of these variants are considered, such as semantic analogy, word formation analogy and the semantics of the borrowed word semantics. Meaning extension is claimed to cause the widespread usage of *po* + dative construction. It can be generalized that 'employees' and 'teachers' tend to govern the latter construction in accordance with the linguistic drift affecting the Russian language since the end of the 19th century. As for 'artists', they prefer the non-prepositional genitive by analogy with the semantically similar word 'modeller'.

EXPERIMENTS IN THE GENERATION OF INFORMATION STRUCTURE OF THE TEXT

Yagunova E. V. (iagounova_elena@mail.ru), St. Petersburg State University, St. Petersburg, Russia

In this paper, an experimental approach is developed to generate information structure of the text. The experimental texts are spoken as opposed to written ones and business texts as opposed to fiction. The texts are analyzed using different experimental settings, viz. (a) soliciting partially predetermined answers from a team of 13 experienced linguists who listened to the spoken texts, (b) soliciting answers from a team of 59 naive subjects who read the written texts, and (c) applying special computer programs to the written texts. The experiments were used to generate a generalized information structure of a text in terms of its constituent utterances viewed as belonging to Topics or Comments. There is an interesting correlation between the classification of the utterances into Topic substructures of the text and the distribution of key words throughout the text. Another result which deserves attention concerns the problem of differentiating between semantic changes induced by specific syntactic shifts and those brought about by information restructuring due to the contextual effects.

EXTRACTING SUBJECTIVE INFORMATION FROM SPOKEN DISCOURSE: THE CASE FOR PROSODIC EMPHASIS

Yanko T. E. (tanya_yanko@list.ru), The Institute for Linguistics, Russian Academy of Science, Moscow, Russia

The aim of this paper is to develop a method for extracting subjective information from spoken discourse. The example of subjective information discussed here is the attitude of the speaker to the reported events as to rare or abnormal. Disrupting a routine course of life which causes the speaker to express his/her strong feelings refers to the concept of emphasis. Emphasis basically has a prosodic expression. The emphatic prosody designates: 'it is great that he did it by himself because generally he needs assistance', 'he came to the meeting which is very unusual because he never comes', 'it was the president who congratulated them which is highly honorable because generally it is the vice-president who does it'. Instances of emphasis are highly frequent in everyday speech as well as in mass media language, especially in breaking news. Reports of achievements and losses, accidents, severe meteorological phenomena, such as hurricanes, earthquakes, swollen streams, and numerical scores of any kind are generally made with the use of emphatic prosody. The instrumental analysis of the physical parameters of emphasis can serve as a means for extracting subjective information from spoken texts. The results presented here are exemplified by records from oral corpora. The frequency tracings are generated by the software program Speech Analyzer.

SYSTEM FOR SEMANTIC ANNOTATION OF DOMAIN-SPECIFIC TEXT CORPORA

Zagorulko M. Ju. (zagulko_maxim@yahoo.com), **Kononenko I. S.** (irina_k@cn.ru), **Sidorova E. A.** (lena@iis.nsk.su), A. P. Ershov Institute of Informatics Systems, Novosibirsk, Russia

A system for universal annotation of text corpus by an expert is presented that contributes to extraction of domain knowledge within the development framework of information systems in specific domains. The technique and software tools for annotation of text corpora allow the expert to carry out two types of semantic annotation: 1) identify text fragments in which the domain concepts represented by special terms actually appear (term annotation) and 2) identify text fragments (often discontinuous) that correspond to domain relations or situations including their participant structure (event annotation). The general principles and schemes of term and event annotation have been formulated and tested for the domain of heterogeneous catalysis on the basis of the hierarchy of term classes chosen in advance. The system, its functional architecture, and user interface are described. Two main directions of usage of semantically annotated texts are discussed: automatic construction of domain lexicons that associate terms with their linguistic and semantic properties; semi-automatic generation of semantic-syntactic patterns for event extraction.

RUSSIAN *RAZOČAROVANIE* IN THE EUROPEAN LINGISTIC CONTEXT: THE PAST AND THE PRESENT

Zalizniak A. A. (anna.zalizniak@gmail.com), Institute of linguistics, Russian Academy of sciences, **Shmelev A. D.** (shmelev.alexei@gmail.com), Moscow Pedagogical State University

The paper discusses the semantic evolution of the word *razočarovanie* (as well as other words of the same family: *razočarovat'*, *razočarovat'sja, razočarovan*) in Russian. It was disclosed that *razočarovanie* being a translation loan-word from the French *désenchantement* had become different from the modern "European" concept encoded in the English *disappointment*, the French *déception*, the German *Enttäuschung*, etc., which refer to the feeling resulting from the failure of expectations. In contrast, the Russian *razočarovanie* refers to the discovery that something highly valued is not as good as one believed it to be. In modern Russian, the meaning of *razočarovanie* and other words of the same family has become less specific and closer to the English *disappointment*; accordingly, their government pattern has been modified. This semantic shift occurred in accordance with the general tendency of loan translation from English, but it is also supported by internal systemic factors. The paper suggests a semantic description of the standard use of the Russian words *razočarovanie*, *razočarovat'(sja)*, *razočarovan* and reveals the semantic shift in question, which is happening before our eyes, but is not readily identified by the speakers of Russian.

IMPROVISED-TEMPORARY-COMPOUNDS AS A NEW EXPRESSIVE MEAN IN RUSSIAN

Zanegina N. N. (zanegina@gmail.com), Russian Language Institute, Russian Academy of Sciences, Moscow, Russian Federation

The article describes improvised temporary compounds, a new means of expressiveness and a pattern of word formation novel to Russian. Such compounds are created in writing by hyphenating a phrase or an entire sentence (cf. we had to invite a woman-who-knew-how-to-mop-the-floor). A corpus-based analysis permitted us to categorize these compound lexemes according to the type of their original syntactic structure (we used a corpus of texts available online in blogs). The following functions of such compounds can be distinguished: 1) logical accentuation, putting an emphasis on a certain meaning which is usually conveyed by several words (car marker lights are usually switched on in twilight, I-haven't-seen-sun-for-days weather); 2) filling in a lexical gap (in some cases, the absence of the sought word can be verbalized: I-do-not-know-how-to-call-that-four-wheeled-thing); 3) isolation of a special class of objects (he normally wears his favorite pants and any-T-shirt-he-found-in-the-closet); 4) citing a direct quote as an example of a typical reaction to an object or an event, or as a telling illustration to an event (he called to say he-was-already-on-his-way-to-the-office); 5) underlining the fact the complement objects depend on this specific main word (me-by-the-window pic taken by Vasya) and some other. We compared hyphenated temporary compounds to such compound lexemes in which 1) all elements are fused (written in one word); 2) all elements are connected by an underscore (_); 3) each element is capitalized. Their functions partially coincide with the functions of the compounds in question.

TOWARDS A SYSTEM OF SYNTACTIC DEPENDENCIES OF GERMAN

Zangenfeind R. (r.zangenfeind@lmu.de), University of Munich, Munich, Germany

A set of syntactic relations in a dependency grammar for German is proposed which can be used for machine translation and other applications. The basis for this set is the formal model of Russian syntax used in the linguistic processor ÈTAP, which is built on the basis of Meaning ⇔ Text theory. Using a terminology as close to ÈTAP as possible facilitates a potential computational implementation. At this stage of development the system of German dependency syntax comprises 58 syntactic relations, which were used to analyze manually several hundred German sentences. 18 of them have definitions that are similar to those of their counterparts in Russian syntax. Seven relations have the same definitions as their Russian counterparts except for the concrete German lexemes that are part of the definitions. For another 30 relations identical definitions as for the corresponding Russian relations can be used. Three German relations do not have a Russian counterpart.

EXPLORING CONTEXT CLUSTERING FOR TERM TRANSLATION

Zhila A. (alisa_zh@mail.ru), **Gelbukh A.** (gelbukh@gelbukh.com) Center for Computing Research, Instituto Politécnico Nacional, Mexico City, Mexico

Many tasks in natural language processing, such as machine translation, word sense disambiguation, word translation disambiguation, require analysis of contextual information. In case of supervised approaches this analysis is performed by human experts, which is very costly. Unsupervised approaches offer fully automatic methods to fulfill these tasks. Yet these methods are not robust, their results are very parameter-dependent and difficult to interpret. Context clustering is an unsupervised technique for analysis of context similarities. In this work we explore dependencies of context clustering results from various clustering parameters. We also explore suitability of the context clustering for word translation disambiguation by evaluating the clustering results against known classes that are classes of translation candidates.

A UNIFIED ANALYSIS OF CLITIC CLUSTERS IN WORLD'S LANGUAGES

Zimmerling A. (meinmat@yahoo.com), Moscow State University for the Humanities, MSUH

The paper proposes a unified analysis of complex syntactic objects defined as clusters. A cluster is by definition a string of elements {a,b,c…n} which can function without combining with each other but are arranged in a rigid order when they assume a contact position, so that for each pair (a, b) the linear order a > b, i.e. 'a precedes b' is fixed. Elements conforming to this definition are called clusterizing. Rules ordering elements in clusters are called Template Rules. In the first section I analyze Template Rules as empiric generalizations made on text corpora representing the normative usage of world's languages from the class of languages with clusterizing clause-level elements. In the final section I analyze Template Rules as linearization algorithms. The general conclusion is that clusters ordered by Template Rules are normally non-homogenous regarding their morpho-syntactic and prosodic values. I furthermore argue that a theory of clusters can be build with little or no resource to the prosody of the clitic elements.

Авторский указатель

Алексеева С. В. ..... т. 1, стр. 51
Анисимович К. В. ..... т. 2, стр. 91
Антонова А. А. ..... т. 2, стр. 104
Антонова А. Ю. ..... т. 1, стр. 616
Апресян В. Ю. ..... т. 1, стр. 1
Архипов А. В. ..... т. 1, стр. 18
Астафьева И. ..... т. 2, стр. 77
Баранов А. Н. ..... т. 1, стр. 28
Беликов В. И. ..... т. 1, стр. 37
Беликова А. Е. ..... т. 1, стр. 187
Богданов А. В. ..... т. 1, стр. 61
Богданова Н. В. ..... т. 1, стр. 71
Большаков И. А. ..... т. 1, стр. 81
Большакова Е. И. ..... т. 1, стр. 81, 490
Бонч-Осмоловская А. А. .... т. 1, стр. 288
Борисова Е. Г. ..... т. 1, стр. 93
Бочаров В. В. ..... т. 1, стр. 51
Браславский П. ... т. 1, стр. 464; т. 2, стр. 1
Будянская Е. ..... т. 1, стр. 296
Валиахметова А. Р. ..... т. 1, стр. 638
Васильев В. Г. ..... т. 2, стр. 66
Васильев П. К. ..... т. 1, стр. 213
Вознесенская М. М. ..... т. 1, стр. 28
Галлямов А. А. ..... т. 1, стр. 502
Гарейшина А. ..... т. 2, стр. 77
Гельбух А. ..... т. 1, стр. 716
Герасименко О. А. ..... т. 1, стр. 162
Гецевич С. А. ..... т. 1, стр. 198
Гецевич Ю. С. ..... т. 1, стр. 198
Грановский Д. В. ..... т. 1, стр. 51
Грачкова М. А. ..... т. 1, стр. 370
Гришина Е. А. ..... т. 1, стр. 173
Гурин Г. Б. ..... т. 1, стр. 187
Давыдов А. Г. ..... т. 1, стр. 122
Давыдов С. ..... т. 2, стр. 66
Даниэль М. А. ..... т. 1, стр. 112
Деликишкина Е. А. ..... т. 1, стр. 129
Добров Г. Б. ..... т. 1, стр. 237
Добровольский Д. О. .... т. 1, стр. 28, 138
Добрушина Н. Р. ..... т. 1, стр. 150
Дрейзис Ю. А. ..... т. 1, стр. 418
Дружкин К. Ю. ..... т. 2, стр. 91
Жила А. ..... т. 1, стр. 716
Загорулько М. Ю. ..... т. 1, стр. 674
Зализняк А. А. ..... т. 1, стр. 684
Занегина Н. Н. ..... т. 1, стр. 696
Захаров Л. М. ..... т. 1, стр. 18
Зеленков Ю. Г. ..... т. 1, стр. 112
Зуев К. А. ..... т. 2, стр. 91
Иомдин Б. Л. ..... т. 1, стр. 213
Иомдин Л. ..... т. 2, стр. 119
Кадыкова А. Г. ..... т. 1, стр. 213
Калинина М. В. ..... т. 2, стр. 51
Кашкин Е. В. ..... т. 1, стр. 227
Кибрик А. А. ..... т. 1, стр. 237
Киселева К. Л. ..... т. 1, стр. 28
Киселева М. Ф. ..... т. 1, стр. 213
Киселёв В. В. ..... т. 1, стр. 122
Клековкина М. В. ..... т. 2, стр. 27
Клыгина Е. А. ..... т. 1, стр. 256
Клюева Н. М. ..... т. 1, стр. 268
Кобозева И. М. ..... т. 1, стр. 277
Кодзасов С. В. ..... т. 1, стр. 18
Козеренко А. Д. ..... т. 1, стр. 28
Кононенко И. С. ..... т. 1, стр. 598, 674
Королева А. ..... т. 2, стр. 77
Корольков К. А. ..... т. 1, стр. 103
Костыркин А. В. ..... т. 1, стр. 288
Котельников Е. В. ..... т. 2, стр. 27
Котов А. ..... т. 1, стр. 296
Кочетков Д. С. ..... т. 1, стр. 122
Кравченко А. Н. ..... т. 1, стр. 319
Краснова Е. В. ..... т. 1, стр. 307
Крейдлин Г. Е. ..... т. 1, стр. 256
Кривнова О. Ф. ..... т. 1, стр. 18
Крылова С. А. ..... т. 1, стр. 331
Крылова Т. В. ..... т. 1, стр. 342
Куниловская М. А. ..... т. 1, стр. 362
Кустова Г. И. ..... т. 1, стр. 352
Кутузов А. Б. ..... т. 1, стр. 362
Кюсева М. В. ..... т. 1, стр. 247
Лебедев А. А. ..... т. 1, стр. 18
Левонтина И. Б. ..... т. 1, стр. 138
Линник А. С. ..... т. 1, стр. 237

Лобанов Б. М. ..... т. 1, стр. 198
Лопухина А. А. ..... т. 1, стр. 213
Лукашевич Н. В. ... т. 1, стр. 490; т. 2, стр. 1
Лукашевич Н. Ю. ..... т. 1, стр. 277
Лучина Е. С. ..... т. 1, стр. 227
Людовик Т. В. ..... т. 1, стр. 383
Люсина В. С. ..... т. 1, стр. 393
Ляшевская О. Н. ... т. 1, стр. 370; т. 2, стр. 78
Малкова А. С. ..... т. 1, стр. 404
Маничева Е. С. ..... т. 1, стр. 418
Матиссен-Рожкова В. И. ..... т. 1, стр. 213
Минлос Ф. Р. ..... т. 2, стр. 91
Мисюрев А. В. ..... т. 2, стр. 104
Митрофанова О. А. ..... т. 1, стр. 370
Михеев М. Ю. ..... т. 1, стр. 431
Мухин М. ..... т. 1, стр. 464
Нехай И. В. ..... т. 1, стр. 477
Нокель М. А. ..... т. 1, стр. 490
Носырев Г. В. ..... т. 1, стр. 213
Орехов Б. В. ..... т. 1, стр. 502
Остапук Н. А. ..... т. 1, стр. 51
Осьмак Н. А. ..... т. 1, стр. 510
Ощепков А. Ю. ..... т. 1, стр. 362
Павлова Е. К. ..... т. 1, стр. 227
Падучева Е. В. ..... т. 1, стр. 522
Пазельская А. Г. ..... т. 1, стр. 616
Панина А. С. ..... т. 1, стр. 288
Петрова М. А. ..... т. 2, стр. 91
Петроченков В. ..... т. 2, стр. 119
Пиперски А. Ч. ..... т. 1, стр. 213
Плешко В. В. ..... т. 2, стр. 51
Поляков А. Е. ..... т. 1, стр. 536
Поляков П. Ю. ..... т. 2, стр. 51
Порошин В. ..... т. 2, стр. 60
Привознов Д. ..... т. 2, стр. 77
Резникова Т. И. ..... т. 1, стр. 227, 288
Рикитянский А. М. ..... т. 1, стр. 213
Романов С. В. ..... т. 1, стр. 370
Рыжова Д. А. ..... т. 1, стр. 247
Савчук С. О. ..... т. 1, стр. 548
Селегей В. П. ... т. 1, стр. 37, 418; т. 2, стр. 91
Семенова С. Ю. ..... т. 1, стр. 568
Сидорова Е. ..... т. 2, стр. 77
Сидорова Е. А. ..... т. 1, стр. 674
Сизов В. ..... т. 2, стр. 119
Смирнова Н. С. ..... т. 1, стр. 307
Соколова Е. Г. ... т. 1, стр. 598; т. 2, стр. 77
Соловьев А. Н. ..... т. 1, стр. 616
Соломенник А. И. ..... т. 1, стр. 607
Степанова М. Е. ..... т. 1, стр. 51
Суриков А. В. ..... т. 1, стр. 51
Ткаченя А. В. ..... т. 1, стр. 122
Толдова С. Ю. ..... т. 2, стр. 77
Тупикина Л. ..... т. 2, стр. 77
Урысон Е. В. ..... т. 1, стр. 627
Федорова О. В. ..... т. 1, стр. 129
Холкина Л. С. ..... т. 1, стр. 247
Худякова М. В. ... т. 1, стр. 237; т. 2, стр. 66
Циммерлинг А. ..... т. 1, стр. 726
Цинман Л. ..... т. 2, стр. 119
Чепуркова А. Ю. ..... т. 1, стр. 362
Четверкин И. И. ... т. 2, стр. 1; т. 2, стр. 15
Чистиков П. Г. ..... т. 1, стр. 103, 607
Шарапов Р. В. ..... т. 1, стр. 578
Шарапова Е. В. ..... т. 1, стр. 578
Шаров С. А. ..... т. 1, стр. 37
Шиморина А. С. ..... т. 1, стр. 370
Шмелев А. Д. ..... т. 1, стр. 587, 684
Шурыгина А. С. ..... т. 1, стр. 370
Ягунова Е. В. ..... т. 1, стр. 652
Янко Т. Е. ..... т. 1, стр. 664
Hein I. ..... т. 1, стр. 443
Kalvik M. ..... т. 1, стр. 443
Kiissel I. ..... т. 1, стр. 443
Mihkla M. ..... т. 1, стр. 443
Moldovan D. ..... т. 1, стр. 454
Pak A. ..... т. 2, стр. 37
Paroubek P. ..... т. 2, стр. 37
Schumann A.-K. ..... т. 1, стр. 559
Sirts R. ..... т. 1, стр. 443
Tamuri K. ..... т. 1, стр. 443
Zangenfeind R. ..... т. 1, стр. 706

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной
Международной конференции «Диалог»

Выпуск 11 (18). 2012

Том 1. Основная программа конференции

Ответственный за выпуск **В. Л. Талис**
Вёрстка **К. А. Климентовский**

Подписано в печать 14.05.2012
Формат 152 × 235
Бумага офсетная
Тираж 200 экз. Заказ № 224

Издательский центр «Российский
государственный гуманитарный университет»
125993, Москва, Миусская пл., д. 6
Тел.: +7 499 973 42 06

Отпечатано с готового оригинал-макета в типографии
ООО «Издательско-полиграфический центр Маска»
117246, Москва, Научный пр-д, д. 20, стр. 9