Отделение историко-филологических наук РАН
Институт русского языка им. В.В. Виноградова РАН
Отдел корпусной лингвистики и лингвистической поэтики

МЕЖДУНАРОДНАЯ НАУЧНАЯ КОНФЕРЕНЦИЯ,

ПОСВЯЩЕННАЯ 20-ЛЕТИЮ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА

Москва, 20-21 декабря 2024 года

Материалы конференции

ББК 81.1
М-43

М-43 **Международная научная конференция**, посвященная 20-летию Национального корпуса русского языка. Москва, 20−21 декабря 2024 г. Материалы конференции. / Отв. ред. С. О. Савчук. − М.: Институт русского языка имени В. В. Виноградова РАН, 2024. –193 с.

Сборник содержит материалы международной научной конференции, посвященной 20-летнему юбилею Национального корпуса русского языка. В центре внимания конференции − лингвистические, программные, технологические аспекты разработки корпусов разных типов и их реализация в НКРЯ, в том числе использование нейросетевых моделей в подготовке и анализе данных. В значительной части работ представлены результаты использования корпусных методов и инструментов в исследовании функционирования и развития русского языка и в практике преподавания.

DOI: 10.31912/nac_corp_20-2024

Содержание

М.И. Берингер, Е.П. Иванова (Санкт-Петербург). Использование национальных корпусов в практике и преподавании перевода…… 8

И.М. Богуславский, В.Г. Диконов, Е.С. Иншакова, А.В. Лазурский, А.А. Мовсесян, С.П. Тимошенко, Т.И. Фролова (Москва). Построение семантически-размеченного корпуса русского языка: SemOntoCor…….. 11

И. А. Бокова (Ижевск). Параллельный корпус НКРЯ: выравнивание, области применения…… 14

Е.О. Борзенко (Москва). Место НКРЯ в методике обучения ..магистранта-русиста научным исследованиям…… 18

П.А. Бычкова (Любляна), *П.В. Падалка, Д.А. Рыжова* (Москва). О методологии корпусных исследований ответных частиц и дискурсивных формул……... 21

Л.А. Велис (Пятигорск). О некоторых затруднениях при использовании национального корпуса русского языка в исследованиях ольфакторной лексики……... 23

Е.Н. Виноградова (Москва) Данные Национального корпуса русского языка при описании функционально-грамматического поля предлога…… 27

Т.А. Гарипов, А.В. Глазкова, Я.Н. Губарькова, А.Д. Козеренко, Д.А. Морозов (Новосибирск, Тюмень, Москва). Упрощение словарных толкований слов русского языка с помощью больших языковых моделей……... 31

А.С. Глаголева (Москва). Возможности Национального корпуса русского языка при исследовании частиц в русском языке……. 34

А.В. Глазкова, Д.А. Морозов, О.А. Митрофанова, С.О. Савчук (Тюмень, Новосибирск, Санкт-Петербург, Москва). Генерация ключевых слов для текстов региональных СМИ с помощью больших языковых моделей…….. 37

Д.А. Девяткин, В.А. Салимовский, Н.В. Чудова (Москва, Пермь) Принцип создания обучающей выборки для большой языковой модели……..… 41

Е.В. Дзюба, О.А. Щербак (Санкт-Петербург). «Служить бы рад...»: НКРЯ — языку для специальных целей…….. 44

Д.О. Добровольский (Москва). Параллельный корпус и двуязычный словарь …… 48

М.В. Дудорова (Екатеринбург), *О.И. Северская* (Москва). Корпусные данные как ключ к фиксации современных грамматических процессов (на материале новых СКС на базе сочетаний с предлогами *в/на/по*)……. 50

А. А. Евдокимова (Москва) Проблемы верификации данных в корпусах древних текстов…….. 54

А. Г. Жукова, О. И. Северская (Москва). Письма «дельные» и «деловые» в эпистолярном наследии А.С. Пушкина и русской лингвокультуре…….. 57

У.С. Загребина (Ижевск). Инструменты Национального корпуса русского языка и их использование в исследованиях лексической сочемаемости прилагательных: диахронический аспект……..….. 61

Н.П. Иордани (Москва). Жанровые характеристики деловых документов в старорусском корпусе…… 65

И.Б. Качинская (Москва). Диалектный корпус: проблемы отбора источников, лингвистической и экстралингвистической разметки…….. 69

Е. В. Кашкин, И. А. Хомченкова (Москва). Круглый стол «Корпусные методы в исследовании языковых контактов»….. 73

Н. И. Киреев (Париж). НКРЯ и историческая акцентология….. 79

О. А. Козан (Анкара). НКРЯ в процессе профессиональной подготовки филолога и переводчика (на примере турецко-русской языковой пары)…… 82

Д. В. Колесова, О. О. Лисова, Т. И. Попова, С. В. Чигинцева (Санкт-Петербург). Языковые маркеры для поиска определенного типа текста в НКРЯ (на примере поиска текстов практического рассуждения)…….. 85

М. В. Копотев (Хельсинки). Корпусная лингвистика и общая теории языка…….. 88

К. М. Корчагин (Москва). Поэтический корпус НКРЯ две декады спустя: результаты и перспективы……... 92

О.Ю. Крючкова, А.И. Буранова (Саратов) Диалектный корпус как ресурс коммуникативной диалектологии……..…. 93

О.Ю. Крючкова (Саратов) Вклад Валентина Евсеевича Гольдина в развитие корпусной диалектологии…… 97

В. П. Лелик, М.Д. Дьячкова, А. С. Сычева, С. В. Дорофеева (Москва), *И. А. Секерина* (Нью-Йорк). Корпус детской речи в формате CHILDES: опыт создания базы данных и применения инструментов компьютерной лингвистики…….. 101

А.Б. Летучий (Москва). Русский/е пассив(ы) с причастием: какие они бывают и как их исследовать?..... 104

О. Н. Ляшевская, С. А. Ребриков (Москва). Лемматизация или дизамбигуация? К проблеме лексико-грамматического анализа сокращений в НКРЯ……. 107

Е.В. Маринова (Нижний Новгород). Варьирование обозначений высокотехнологичных человекоподобных программ в отношении одушевлённости/неодушевлённости: условия и причины……...… 110

Е.В. Маркасова (Пекин). «Моё Я» в русском языке XIX-XX вв. по данным НКРЯ……..… 114

Д.А. Морозов, А.В. Глазкова, Я.Н. Губарькова, Т.А. Гарипов, С.С. Столяров, Н.А. Власова, О.Н. Ляшевская, И.А. Смаль, А.Д. Козеренко (Новосибирск, Тюмень, Москва). Применение инструментов обработки естественного языка на базе машинного обучения при разработке корпусов: опыт Национального корпуса русского языка……..……. 118

Ю.В. Николаева (Москва). Возможные направления работы в организации и аннотации мультимодальных корпусных данных……. 122

С.И. Переверзева (Москва). Глубокая разметка жестов в мультимедийном русском корпусе (МУРКО): проблемы и перспективы (2016–2024)……. 125

З.Ю. Петрова, Н.А. Фатеева (Москва). Использование Национального корпуса русского языка при словарном описании системы метафор и сравнений русской литературы в ее динамике…… 129

В.И. Подлесская (Москва). МУРКО и его братья: звуковые корпуса в исследованиях устной речи……..…. 133

Поляков А.Е. (Москва). Лемматизаторы для национального корпуса русского языка……... 136

Н. А. Ребецкая (Москва). База данных словаря языка Пушкина. Корпусные исследования……. 142

Л. В. Рычкова (Гродно). НКРЯ как источник для формирования экспериментально-доказательной базы научных исследований магистрантов-лингвистов…… 148

Сабольч Янурик (Будапешт). Вопросы применения корпусных данных в лексикографическом описании английских заимствований в русском языке……..… 152

И. А. Смаль, Д. А. Морозов (Новосибирск). Разметка семантики в системе НКРЯ……..…. 153

Т. П. Соколова (Москва). Возможности использования Национального корпуса русского языка для производства судебной лингвистической экспертизы……. 156

В. Д. Соловьев (Казань). Полнота и сбалансированность: сопоставительный анализ НКРЯ и Google Books Ngram…….…. 160

И. И. Столяров, О. А. Митрофанова (Санкт-Петербург). Автоматическое разграничение омографов (на материале НКРЯ)…….…. 163

Е.В. Туркина (Москва). Искусственный интеллект в медийных и сетевых текстах: черты корпусного портрета……...…. 166

Е.В.Филиппова (Москва). Трансформация смыслового наполнения понятия «милость»…….…… 170

М. В. Хохлова (Санкт-Петербург). Поисковые интерфейсы в корпусах текстах: сравнение возможностей и ограничений…. 174

Л. Л. Шестакова, А. С. Кулева (Москва). Использование ресурсов Национального корпуса русского языка в работе над «Словарем языка русской поэзии XX века»…….... 177

К.М. Шилихина (Воронеж). Корпусные данные в исследованиях метакоммуникации……..… 181

Р. И. Шмурак (Ханчжоу, Китай). Нейронные сети как инструмент корпусного поиска…… 185

К.А. Щукина (Санкт-Петербург). К вопросу об использовании Национального корпуса русского языка в выпускных квалификационных работах……..... 189

М.И. Берингер, Е.П. Иванова
(Санкт-Петербург, Россия)
Санкт-Петербургский государственный университет
m.beringer@spbu.ru, e.ivanova@spbu.ru

ИСПОЛЬЗОВАНИЕ НАЦИОНАЛЬНЫХ КОРПУСОВ В ПРАКТИКЕ И ПРЕПОДАВАНИИ ПЕРЕВОДА

Рассматриваются возможности использования данных национальных языковых корпусов при переводе художественных текстов предшествующих эпох. Особое внимание уделяется проблеме отбора лексем с учетом времени их появления в языке перевода.
Ключевые слова: национальный корпус; художественный перевод; архаизация; диахронический фактор в переводе

Появление национальных языковых корпусов открывает новые возможности для практики и преподавания перевода, а также для дальнейшего развития исследований в рамках переводоведения [Семина 2014: 336-341]. В этой работе мы остановимся лишь на одном аспекте их использования, а именно целесообразности обращения к информационным ресурсам такого типа при переводе художественных текстов предшествующих эпох.

Перевод текстов, созданных во временные периоды, более или менее удаленные от современности, ставит перед переводчиком целый ряд задач. К классической переводческой проблеме конфликта взаимодействия двух национальных культур добавляется проблема изменения языка во времени, которая в свою очередь делится на две другие составляющие: архаичность языка оригинала и необходимость поиска переводческих решений, учитывающих этот исторический аспект в языке перевода. Иными словами, на этапе чтения оригинала переводчику следует понять значение слова или фраземы в определенный исторический срез, а в процессе перевода — создать такой текст, который, с одной стороны, будет нести в себе дух соответствующей эпохи, но при этом не будет архаизирован до такой степени, чтобы стать трудным для восприятия современным читателем.

Переводчики, работающие со старыми текстами, осознают, что, переводя текст, написанный в момент *t*, на самом деле они переводят текст, который они *читают в момент t'* [Vrinat-Nikolov 2003: 66-76]. Объективная невозможность воспринимать текст так, как его воспринимал бы современник автора, может при этом усугубляться

непониманием отдельных элементов текста и приводить к ошибкам в переводе. Это тем более вероятно в тех случаях, когда переводимая единица не представлена в словарных базах, или контекст ее употребления не позволяет определить ее значение, или данная единица не встречалась переводчику в других эксплицирующих контекстах с достаточной степенью репрезентативности. До появления национальных корпусов и создания лексикографических баз данных, учитывающих диахронические аспекты семантики слов, при решении таких задач полагались исключительно на эрудицию, филологическую культуру и интуицию переводчика.

Появление национальных языковых корпусов позволяет решить эту проблему благодаря возможностям, которые предлагает аннотирование текстов, в частности поиска с учетом метаданных и лексической, морфологической и синтаксической разметки: время и место создания текста, его жанровая характеристика, диалектная специфика и т.д. Получение доступа к множеству контекстов, относящихся к определенному этапу развития языка, помогает переводчику понять не только значение лексемы или выражения, но также и оценить ее место в определенной семиотической системе.

Однако на втором этапе перевода, а именно на этапе *создания* переводного текста, доступ к ресурсам национальных языковых корпусов представляется еще более важным для положительного результата в работе переводчика.

Даже если бы переводчик владел формой бытования языка определенного хронологического этапа — например, русским языком петровской эпохи, едва ли было бы целесообразно использовать этот язык для перевода текста, созданного на рубеже XVII- XVIII вв. в другой языковой культуре. Стремясь передать образ определенной эпохи через выразительные средства переводящего языка, переводчик часто прибегает к приему архаизации. Обычно он делает это не за счет использования всей совокупности языковых средств, а прибегает к использованию отдельных маркеров — как правило лексических. Средства архаизации позволяют искусственно «состарить» язык перевода. Однако не только специалисты, но даже неискушенные читатели замечают, что применительно к разным эпохам переводчики нередко применяют одинаковые средства архаизации, т.е. стилизация под эпоху конца XVIII века может мало отличаться от стилизации под язык более ранних или более поздних эпох.

В настоящее время, благодаря доступу к национальным языковым корпусам, переводчик может не только узнать время появления

той или иной лексемы в языке перевода, но и проследить эволюцию ее значения. Это позволяет ему избежать избыточной архаизации текста перевода и, наоборот, не допустить использования в тексте перевода лексем, еще не существовавших в языке перевода в соответствующую эпоху. Например, в переводах на русский язык французских текстов первой половины XVIII века нами были обнаружены такие лексические единицы как *комментатор, портфель, расстройство желудка* и т.п., которые, по данным НКРЯ, вошли в русский язык только в начале XIX века. Там же встретилось слово *администрация* в значении «служба, ведомство», тогда как, судя по данным, полученным из НКРЯ, в первой половине XVIII века оно использовалось только в значении «процесс управления». Странным показалось также присутствие в одном из переводов текстов этой эпохи междометия *здорово* для выражения одобрения. Полагаем, что если бы переводчик обратился к данным НКРЯ, едва ли он стал бы употреблять это слово в данной функции, закрепившейся за этой лексемой только в XIX веке.

Таким образом, информация, которую переводчик может почерпнуть при обращении к корпусным данным, помогает ему найти оптимальное решение при выборе переводческих эквивалентов. Использование приема архаизации через употребление лексем, существовавших в языке перевода на соответствующем исходному тексту историческом этапе, открывает перед переводчиком возможность достижения высокого уровня художественного перевода, основанного на воссоздании равнозначного исторического контекста двух текстов.

Литература

Семина О.Ю. Об использовании данных национального языкового корпуса при переводе // Известия ТулГУ. Гуманитарные науки. 2014. №1. С. 336–341.

Vrinat-Nikolov M. Mais que traduit-on quand on traduit ? // Ouvrir les archives « Henri Meschonnic », sous la dir. de Serge Martin. Mont de Laval, L'Atelier du Grand Tétras, 2013. P. 66–76.

Национальный корпус русского языка. URL: http://ruscorpora.ru

И.М. Богуславский, В.Г. Диконов, Е.С. Иншакова, А.В. Лазурский, А.А. Мовсесян, С.П. Тимошенко, Т.И. Фролова
(Москва, Россия)
Институт проблем передачи информации им. А.А.Харкевича
bogus@iitp.ru, sdiconov@mail.ru, e.s.inshakova@gmail.com, lazursky@mail.ru, andrey.movsesyan@frtk.ru, timoshenko@iitp.ru, tfrolova@gmail.com

ПОСТРОЕНИЕ СЕМАНТИЧЕСКИ-РАЗМЕЧЕННОГО КОРПУСА РУССКОГО ЯЗЫКА: SemOntoCor[1]

Семантически размеченный корпус SemOntoCor представляет собой следующий шаг в развитии корпуса SynTagRus, добавляя к существующим в нем видам разметки Базовые семантические структуры (БСемС). БСемС охватывает различные аспекты семантики — предикатно-аргументную структуру, лексическую семантику, семантические роли, темпоральную семантику, синтаксическую семантику, анафору, модальность. Первая очередь SemOntoCor представляет собой размеченный русский перевод повести-сказки Антуана де Сент-Экзюпери «Маленький принц» (1532 предложения, 13120 токенов). В настоящее время ведется разметка второй очереди корпуса, которая включает тексты из корпуса SynTagRus общим объемом свыше 10 000 предложений.

Ключевые слова: семантический корпус, семантическая разметка, онтология, СинТагРус, Базовая семантическая структура, инструментарий пользователя.

В Институте проблем передачи информации РАН разрабатывается семантический корпус русского языка, основанный на онтологии (SemOntoCor). Подробнее о нем см. в [Boguslavsky et al. 2023].

SemOntoCor имеет несколько целей.

1. SemOntoCor как представительная коллекция семантических структур русского языка, позволяющая изучать взаимодействие лексики, синтаксиса и семантики на реальных текстах. В этом отношении SemOntoCor следует в русле НКРЯ, содержащего целую серию аннотированных подкорпусов, которые демонстрируют разнообразные срезы русского языка, дополняющие друг друга.

[1] Работа выполняется в рамках гранта РНФ 24-18-00988.

2. SemOntoCor как датасет для машинного обучения. Во всем мире имеется относительно немного семантических датасетов широкого профиля, особенно в сравнении с датасетами в области синтаксиса. Это особенно справедливо, если речь идет о русском языке. Для русского языка уже давно существуют качественные корпуса текстов, размеченные синтаксическими структурами, которые предоставляют обширный материал для машинного обучения, затрагивающего лексический и морфо-синтаксический уровень языка. Для уровня семантики таких корпусов для русского языка, к сожалению, практически нет. SemOntoCor станет первым корпусом такого рода для русского языка.

3. SemOntoCor как ресурс, тесно скоррелированный с правиловым семантическим анализатором SemETAP, нацеленным на автоматическое построение семантических структур и извлечение из них разного рода следствий. Формализм и принципы построения семантических структур, принятые в SemOntoCor, непосредственно унаследованы из проекта SemETAP. Поэтому SemOntoCor и SemETAP могут естественным образом обогащать друг друга: SemETAP может быть использован для разметки SemOntoCor в полуавтоматическом режиме, а SemOntoCor может применяться для регрессионного тестирования SemETAP.

SemOntoCor обладает следующими особенностями.

1. SemOntoCor представляет собой следующий шаг в развитии корпуса SynTagRus, добавляя к существующим в нем видам разметки (морфологической, синтаксической, микросинтаксической, лексико-функциональной, кореферентной и нек. др.) семантические структуры. Благодаря этому появляется возможность на едином теоретическом базисе получить ресурс, имеющий несколько слоев разметки, начиная от морфологической структуры и кончая семантической.

2. SemOntoCor разрабатывается в связке с модулем семантического анализа (SemETAP), который вводит в лингвистический процессор ЕТАР семантический уровень [Богуславский 2021]. На этом уровне предложение представляется двумя структурами — Базовой семантической структурой (БСемС) и Расширенной семантической структурой (РСемС). Первая из них отражает непосредственное значение предложения, а вторая обогащает первую разнообразными следствиями. Оба типа структур строятся из элементов онтологии — особого ресурса, который представляет собой структурированное представление объектов внешнего мира и их свойств. Семантические

структуры записываются на специально разработанном языке Etalog [Rygaev 2018], удобном как для представления семантической структуры, так и для формулировки правил вывода следствий.

3. Корпус SemOntoCor сопоставляет каждому предложению его БСемС. БСемС охватывает различные аспекты семантики — предикатно-аргументную структуру, лексическую семантику (частично), семантические роли, темпоральную семантику, синтаксическую семантику, анафору, модальность. В ряде случаев осуществляется разложение лексического значения на более мелкие семантические элементы. Лексические и грамматические значения представляются с помощью одних и тех же семантических элементов.

4. Разметка корпуса может производиться вручную или полуавтоматически. В последнем случае БСемС строится посредством анализатора SemETAP, а затем редактируется экспертом.

5. Интерфейс пользователя позволяет осуществлять поиск в корпусе по русским словам, по семантическим элементам, по семантическим ролям и по произвольному выражению на языке Etalog. БСемС можно визуализировать в виде графа. В нем наглядно видна не только сама семантическая структура, но и связи между фрагментами БСемС и соответствующими им словами предложения.

6. Интерфейс аннотатора позволяет осуществлять ручную разметку, а также запускать для разметки анализатор SemETAP и затем редактировать его результат в графическом редакторе.

Первая очередь SemOntoCor представляет собой разметку русского перевода повести-сказки Антуана де Сент-Экзюпери «Маленький принц» (1532 предложения, 13120 токенов). В настоящее время ведется разметка второй очереди корпуса, которая включает тексты SynTagRus-а объемом свыше 10 000 предложений.

Литература

Богуславский И.М. Семантический анализ с опорой на умозаключения в функциональной модели языка. // Вопросы языкознания, N1, 2021, с. 29–56.

Boguslavsky I., V. Dikonov, T. Frolova, E. Inshakova, L. Iomdin, A. Lazursky, I. Rygaev, S. Timoshenko. Constructing a Semantic Corpus for Russian: SemOntoCor. // Computational linguistics and intellectual technologies. Proceedings of the International Conference "Dialogue 2023".

Rygaev I. Etalog-a natural-looking knowledge representation formalism // ИТиС 2018, 473–482.

И. А. Бокова
(Ижевск, Россия)
ФГБОУ ВО «ИжГТУ имени М. Т. Калашникова»
ibokova899@gmail.com

ПАРАЛЛЕЛЬНЫЙ КОРПУС НКРЯ: ВЫРАВНИВАНИЕ, ОБЛАСТИ ПРИМЕНЕНИЯ

В работе продемонстрированы возможности применения инструментов переводного корпуса при исследовании прямой и косвенной речи в английском параллельном корпусе НКРЯ, предпринята попытка определить потенциальные переводные лексические эквиваленты, а также показано применение параллельного корпуса при диахронических исследованиях лексики и грамматики.

Делаются выводы о том, что сопоставление лексических и грамматических особенностей оригинального текста и текста, содержащего его перевод, позволяет получить лингвистически обоснованные результаты.

Ключевые слова: параллельный корпус, выравнивание, перевод, НКРЯ, HunAlign

Целью данной работы является изучение и анализ инструментов параллельного корпуса НКРЯ, а также описание возможностей применения его в лингвистических исследованиях. В рамках настоящего исследования были поставлены задачи рассмотреть возможности применения результатов изучения и анализа данных двуязычного корпуса в различных научных областях.

1. Параллельный корпус представляет собой сопоставление текста языка-источника и его перевода на язык-цель, представленное в выровненном виде [Баранов 2001: 63]. Чёткое соответствие между аналогичными фрагментами текста позволяет сразу же определить сходства или различия текста и его перевода.

Программа HunAlign, используемая в НКРЯ, производит выравнивание двуязычного текста на уровне предложений. В качестве входных данных применяется двуязычный текст, токенизированный и разделённый на предложения. Выходными данными является последовательность двуязычных пар предложений (bisentences).

При наличии переводного словаря программа использует его данные, комбинируя их с данными о длине предложений в соответствии с алгоритмом выравнивания Гейла-Черча.

В случае отсутствия словаря HunAlign сначала возвращается к данным о длине предложения, а затем создаёт собственный автоматический словарь на основе первичного выравнивания. При втором проходе алгоритмы, уже используя составленный автоматический словарь, заново выравнивают текст [HunAlign].

2. Результаты исследований, проведенных при помощи параллельного корпуса НКРЯ, в котором на 2024 год доступно 32 языка, применяются в таких прикладных областях, как типология, контрастивная лингвистика, внутриязыковое варьирование, теория и практика перевода, обучение языкам и машинное обучение. К сфере машинного обучения относится также обучение SMT и автоматический контроль верности перевода [Волченкова 2015: 34].

Например, говоря о переводе, в английском параллельном корпусе НКРЯ была проведена попытка определить потенциальные переводные эквиваленты леммы *scowl*. Среди 393 вхождений выделены варианты перевода *набычился, сердито посмотрел, угрюмая гримаса, хмурится, сердито сверкнуть глазами, хмуро глянул,* наиболее частотными из которых стали *хмуриться, нахмуриться*, в то время как остальные результаты представляют собой лишь стилистические варианты с различной степенью проявления признака. Однако относительно параллельного корпуса НКРЯ говорить о стилистической окраске слов довольно сложно, поскольку среди дополнительных меток присутствует лишь признак “словарное” и “несловарное” (*norm, bastard*).

На примере исследования прямой и косвенной речи в английском параллельном корпусе НКРЯ были рассмотрены некоторые трансформации текста, возникшие при переводе в связи с синтаксическими особенностями языка-оригинала и языка-перевода.

(1) *Как досадно, что он сказал матери, что пойдет. (Успенский). / What a nuisance having told his mother that he would go. (Ouspensky).*

В данном примере при переводе соединение главной и придаточной части предложения осуществляется за счёт полупреди-

кативного обстоятельства *having told*, «сказав», тогда как в оригинале использовалось придаточное предложение.

(2) *“Well, it's not very easy to describe, is it, Edmund?” said the High King.(Льюис). / — Ну, это нелегко описать, — ответил Верховный Король, — да, Эдмунд? (Островская).*

Авторская ремарка в данном примере находится в постпозиции в оригинале, а в переводе — в интерпозиции, что обусловлено наличием разделительного вопроса (англ. *Tag Question*). Последнюю часть разделительного вопроса, *is it*, обычно переводят как «не так ли», однако поскольку для данного произведения осуществлялся художественный перевод, то буквальное значение так называемого «хвостика» передаётся с помощью «*да*».

(3) *“Oh. Thanks,” said Richard. He couldn't think of anything to say. (Гейтман). / — Это хорошо, — обрадовался Ричард и тут же обнаружил, что не знает, о чем её спросить. (Конча, Мельниченко).*

Используемый в этом примере приём объединения предложений применяется для достижения грамматической и стилистической адекватности перевода, а также для устранения потенциальной недостаточности смысла как самих отдельных предложений, так и связи между ними.

Так, в связи с отсутствием определенных синтаксических конструкций в одном из сравниваемых языков текст трансформируется в целях сохранения семантической эквивалентности и адекватности перевода.

Параллельный корпус НКРЯ может использоваться в области диахронических исследований лексики и грамматики. Запросная форма *тебе* для современных текстов выдаёт результат *you*, а для более старых текстов, до середины XIX века — *thee*, устаревшую форму косвенного падежа местоимения *thou* (совр. *you*). Следует отметить, что в современных текстах словоформа *thee* также встречается, однако имеет больше стилистическую функцию, передающую особенности речи персонажей художественных произведений или авторской стилизации текста, а не непосредственного употребления слова в соответствии с языковой нормой.

Таким образом, соотношение фрагментов текста оригинала и перевода на уровне предложений в параллельном корпусе НКРЯ дает

возможность сопоставить варианты перевода конкретных слов, сравнить синтаксические конструкции в разных языках и проследить изменения, происходящие с лексикой с течением времени.

Литература

Баранов А. Н. Введение в прикладную лингвистику: Учебное пособие. — М.: Эдиториал УРСС, 2001. 360 с.

Волченкова К. Н. Параллельный корпус как справочная база данных в работе переводчика // Проблемы и перспективы развития образования в России, 2015. №33. С. 32–35.

НКРЯ — Национальный корпус русского языка [Электронный ресурс]. URL: https://ruscorpora.ru (дата обращения: 27.10.2024).

HunAlign. Sentence aligner [Электронный ресурс]. URL: http://mokk.bme.hu/resources/hunalign/ (дата обращения: 26.10.2024).

Е.О. Борзенко
(Москва, Россия)
Православный Свято-Тихоновский гуманитарный университет (ПСТГУ)
ekborzenko@yandex.ru

МЕСТО НКРЯ В МЕТОДИКЕ ОБУЧЕНИЯ МАГИСТРАНТА-РУСИСТА НАУЧНЫМ ИССЛЕДОВАНИЯМ

В докладе представлен опыт преподавания дисциплин, связанных с академическим письмом, на материале НКРЯ, ГИКРЯ и «естественных корпусов» Яндекс и Гугл для магистров первого курса, обучающихся по программе «Русская и славянская филология». Автор описывает методику проектной работы с учетом того, что по крайней мере некоторые из учащихся плохо знакомы или не знакомы вовсе с лингвистическими корпусами, и обосновывает актуальность использования неологизмов как предмета корпусного анализа для обучения начинающих исследователей.

Ключевые слова: корпус, корпусная лингвистика, НКРЯ, ГИКРЯ, методика обучения русистов, обучение научному исследованию, академическое письмо

Одной из особенностей обучения в современной магистратуре является возможность перейти в нее после окончания бакалавриата любой другой специальности. В связи с этим оказывается, что магистры первого курса направления, изучающего русскую филологию, порой даже не знают о существовании НКРЯ. Впрочем, выпускники-филологи также далеко не всегда свободно работают с этим инструментом. Как же максимально быстро ввести студентов магистратуры в методику работы с лингвистическими корпусами?

На наш взгляд, целесообразно реализовать эту задачу в форме проектной работы с осязаемым результатом, а именно: в течение семестра каждый студент должен создать тезисы для научной конференции или опубликовать статью, которые обычно делаются на основе корпусных данных и служат формой отчётности в семестре. Такая работа проводится на занятиях по дисциплине «Академическое письмо». Студенты, уже защитившие ВКР, хорошо знакомы с основными особенностями научного стиля, и написание научной статьи и апробация их работы становится позитивным фактором в их становлении как исследователей.

Таким образом, в начале семестра преподаватель предлагает выбрать каждому свое слово / форму / конструкцию, которую студент будет изучать в течение семестра. С нашей точки зрения, в этом контексте особенно «выгодно» выбирать неологизмы. Во-первых, они обычно интересны учащимся, а во-вторых — мало описаны, что даёт дополнительные преимущества: студент не должен изучать большой объем литературы (что значимо, особенно если тема его будущей магистерской не связана с корпусной лингвистикой). Кроме того, для неологизмов обычно существует не так много примеров, по крайней мере в НКРЯ, почему меньше риск «потонуть» в огромном количестве результатов поиска. И наконец, именно изучая неологизмы, можно учиться не только микродиахроническому, но и нанодиахроническому корпусному анализу.

Первый этап — общее знакомство с НКРЯ, в контексте новых слов — с Основным корпусом, Газетным корпусом и с корпусом «Социальные сети». В частности, студенты сразу осваивают форму лексико-грамматического поиска и поле «Лемма» для лексических неологизмов, «Словоформа» — для новых форм и функцию «+ слово» — для новых конструкций. Из других полезных элементов можно отметить «График» и «Портрет слова», хотя обе эти функции не всегда показательны для новых слов. Сразу же осваивается возможность сортировки примеров по году создания текста и выгрузки примеров: хотя в НКРЯ очень удобные возможности для копирования результатов, гораздо продуктивнее работать со скачанными примерами в Excel, где студенты могут визуально выделить (например, цветом) примеры с повторяющимися значениями и особенностями употребления. Если файл изначально отсортирован по году написания текстов, то появляется возможность проследить появление новых значений и особенностей употребления в диахронии, а также не потерять интересные примеры.

Второй этап — освоение ГИКРЯ. Для начала студенты проводят первичный поиск слова (с настройкой выдачи на 100 результатов, чтобы понять, как настроить поисковый запрос, а ещё важнее — насколько слово в принципе перспективно для исследования). После этого студенты ставят в настройку выдачи большое число (например, 10 млн. примеров), скачивают выгрузку и работают с ней по тому же принципу, что и с выгрузкой из НКРЯ. Также используется выгрузка «Год написания», в некоторых случаях — «Дата рождения». Результат

исследования — все значимое, начиная от момента появления слова, системы его значений и кончая сочетаемостью, — оформляется в виде текста с подзаголовками, куда встраивается и анализ из НКРЯ.

Третий этап — возвращение к НКРЯ. После того как студенты выстроили для себя представление о неологизме, они могут сказать, каковы его синонимы. В этом могут помочь не только данные словарей и результаты анализа, но и неформальные обсуждения этого неологизма в Рунете (дискуссии типа «Зачем в русском языке слово рандомный»), где порой встречаются очень интересные интуитивные находки. Составив для себя список основных синонимов для каждого значения слова, студент возвращается в Основной корпус НКРЯ, настраивает выдачу примеров для каждого синонима и просматривает, например, первые 30 примеров, в каждом пытаясь устно заменить синоним на свой неологизм. В большинстве случаев это невозможно по стилистическим причинам, что также отмечается студентом, однако его основная задача — найти примеры, где замена невозможна не только по причине неформального характера языковой единицы, но и по семантическим, синтаксическим и другим причинам. Примеры, в которых замена невозможна, позволяют конкретизировать значение и особенности употребления исследуемого неологизма, но самое главное — сделать масштабный вывод о том, зачем языку нужна эта единица, какую именно лакуну она заполняет.

Также ведется работа с «естественными корпусами» Яндекс и Гугл, которые приходится использовать для поиска самых новых слов и для анализа положения слова в текущем году, а также для поиска самых ранних примеров. Возможна тренировка навыков поиска в ВК, Телеграм и т.д., однако это происходит в тех случаях, когда остальные опции не достаточны.

Таким образом, благодаря работе с НКРЯ и ГИКРЯ студенты приобретают навыки научной работы, и если в контексте неологизмов ГИКРЯ зачастую более важен, то НКРЯ остается той базой, на основе которой начинается корпусное исследование. В некоторых случаях НКРЯ даже более «выгоден»: например, при поиске единиц, которые встречаются в диалогах (*дап, еп*) и относительно привычных слов типа *спойлер*. Иногда проектные работы на основе корпуса с последующим выступлением на конференции и публикацией статьи служат основой для интересной магистерской диссертации.

[1]П.А. Бычкова, [2]П.В. Падалка, [2]Д.А. Рыжова
[1](Любляна, Словения)
Университет Любляны
[2](Москва, Россия)
Национальный исследовательский университет
«Высшая школа экономики»
[1]polyatomson@gmail.com, [2]pvpadalka_1@edu.hse.ru,
[2]daria.ryzhova@mail.ru

О МЕТОДОЛОГИИ КОРПУСНЫХ ИССЛЕДОВАНИЙ ОТВЕТНЫХ ЧАСТИЦ И ДИСКУРСИВНЫХ ФОРМУЛ

Доклад посвящен особому типу лингвистических единиц — устойчивым ответным репликам, характерным для жанра бытового диалога. Мы покажем, какими особенностями обладают такие единицы, и обсудим, какого рода тексты и аннотация необходимы для полноценных и плодотворных корпусных исследований подобных явлений.

Ключевые слова: диалог, ответные реплики, дискурсивные формулы, корпусные исследования, речевые акты

Ответные частицы (прежде всего, *да* и *нет*), а также выполняющие аналогичные прагматические функции дискурсивные формулы (*Ещё чего! А как же!* и др., см. [Рахилина и др. 2021]) до сих пор недостаточно исследованы ни в русском языке, ни, тем более, в типологической перспективе. Одна из основных причин появления этой лакуны — привязка этих лингвистических единиц к очень специальному речевому жанру: прежде всего, жанру устного бытового диалога.

С одной стороны, именно этот жанр можно считать в каком-то смысле самым базовым: именно в нем проявляется основное, коммуникативное, предназначение языка, именно этот функциональный стиль доступен в первую очередь любому живому языку, в том числе миноритарному, не имеющему ни письменности, ни государственного статуса. Казалось бы, как раз этот жанр и должен был бы изучаться лингвистами прежде всего.

С другой стороны, этому жанру обычно уделяется значительно меньше внимания, чем письменным текстам или даже устным нарративам. Происходит это по очень простой причине: устный бытовой диалог сложнее всего задокументировать. Эта особенность ярко иллюстрируется Национальным корпусом русского языка,

который по праву считается образцовым примером сбалансированного и репрезентативного национального корпуса. НКРЯ содержит огромное количество тщательно размеченных устных и письменных текстов самых разных жанров и временных периодов, от классической художественной литературы до постов и комментариев в социальных сетях, от XI века до самых современных употреблений. Тем не менее, жанр устного бытового диалога очень ограниченно представлен и здесь: такие фрагменты встречаются в текстах пьес в основном подкорпусе, в некоторых художественных фильмах в мультимедийном подкорпусе и, конечно, в устном подкорпусе, объем которого пока не очень велик. Кроме того, существенная часть диалогов в устном подкорпусе — это интервью, а не бытовое коммуникативное взаимодействие между собеседниками. Ответные частицы и дискурсивные формулы очень чувствительны к типу речевого акта стимульной реплики, поэтому для их исследования важны диалоги разных типов: в интервью в качестве стимульных реплик выступают, как правило, вопросы и предположения, а разного рода директивы (просьбы, требования, просьбы разрешить) характерны скорее для бытового общения.

Другое обстоятельство, затрудняющее исследование прагматических выражений в диалоге на материале корпусов, — это технические возможности поиска. Так, например, для анализа диалога такая структурная единица, как реплика, во многих случаях важнее, чем предложение. Между тем, насколько нам известно, в НКРЯ нет возможности поиска реплики целиком, той или иной единицы в начале или в конце реплики, а также возможности скачивания выдачи с сохранением левого и / или правого контекста в терминах реплик, например, дискурсивной формулы вместе с репликой-стимулом.

В докладе мы расскажем об особенностях дискурсивных формул как таковых и порассуждаем о том, какого рода тексты, разметка и возможности поиска позволили бы исследовать их более полно и продуктивно.

Литература

Рахилина, Е. В., Бычкова, П. А., Жукова, С. Ю. Речевые акты как лингвистическая категория: дискурсивные формулы // Вопросы языкознания. 2021. № 2. С. 7–27.

Л.А. Велис
(Пятигорск, Россия)
Пятигорский государственный университет
lolla-99@mail.ru

О НЕКОТОРЫХ ЗАТРУДНЕНИЯХ ПРИ ИСПОЛЬЗОВАНИИ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА В ИССЛЕДОВАНИЯХ ОЛЬФАКТОРНОЙ ЛЕКСИКИ[1]

В данной работе рассматривается использование Национального корпуса русского языка (НКРЯ) для лингвистических исследований на примере ольфакторной (запаховой) лексики. Обсуждаются преимущества и недостатки применения корпуса, а также технические проблемы, возникающие при обработке языкового материала. Особое внимание уделяется инструменту «портрет слова».

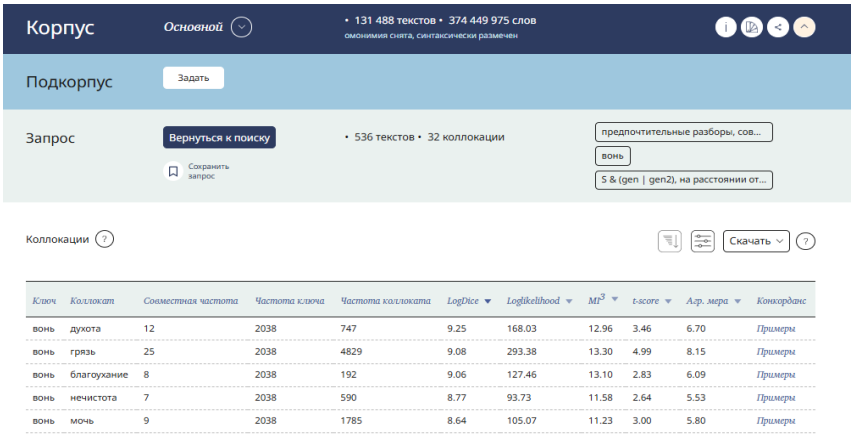
Ключевые слова: Национальный корпус русского языка, ольфакторная лексика, квантитативная лингвистика, портрет слова.

Обширность баз данных и технического функционала языковых корпусов сегодня позволяет говорить о том, что корпус становится удобным инструментом для осуществления сравнительно быстрых и количественно обширных исследований в рамках квантитативной лингвистики [Павельева 2016, Палийчук 2020].

Результаты, получаемые из больших объемов данных, призваны показать тенденции и зависимости, не выявляемые на материалах меньшего объема. Например, контекстуальное окружение ольфакторной лексики («запах», «аромат» и т.д.) позволило бы говорить об ольфакторной картине в русском и других языках.

В НКРЯ данный анализ возможен благодаря инструменту «портрет слова», в котором более тонкие настройки позволяют находить слова по их грамматическим функциям с помощью поиска по коллокатам. Данные запроса структурированы в виде таблиц с помощью различных метрик (LogDice, t-score, MI3, Loglikelihood) (рис.1).

[1] Исследование выполнено за счет гранта Российского научного фонда № 24-28-00540, https://rscf.ru/project/24-28-00540/



Ключ	Коллокат	Совместная частота	Частота ключа	Частота коллоката	LogDice	Loglikelihood	MI³	t-score	Агр. мера	Конкорданс
вонь	духота	12	2038	747	9.25	168.03	12.96	3.46	6.70	Примеры
вонь	грязь	25	2038	4829	9.08	293.38	13.30	4.99	8.15	Примеры
вонь	благоухание	8	2038	192	9.06	127.46	13.10	2.83	6.09	Примеры
вонь	нечистота	7	2038	590	8.77	93.73	11.58	2.64	5.53	Примеры
вонь	мочь	9	2038	1785	8.64	105.07	11.23	3.00	5.80	Примеры

Рис. 1. Выборка коллокатов к существительному «смрад»

Однако вместе с тем выборка вызывает ряд затруднений, связанных, во-первых, с особенностями языкового материала, и, во-вторых, с технологиями, лежащими в основе обработки текста.

Первое затруднение само по себе не является полностью непреодолимым, т.к. НКРЯ предоставляет количественные данные, и пользователь может вручную просмотреть все примеры. Однако при исследовании примеров видно, что уже на этом уровне появляются некоторые «зашумления» в виде дубликатов предложений. Эта особенность возникает из-за того, что корпус допускает наличие одного и того же материала, изданного в разные годы. Возможно, для статистического анализа ольфакторной лексики данные погрешности не являются критическими, поскольку тема «запахов» в принципе представлена не так обширно. Но вопрос остается открытым: насколько повлияет данная «зашумленность» при исследовании в других областях? Более того, может ли исследователь быть уверен в том, что в базе, содержащей 131 488 текстов, искомое слово встречается всего лишь в 536 из них?

Второе затруднение принципиально важно, потому что при просмотре примеров, достаточно часто можно видеть, что корпус:

А) выводит ошибочные данные (на рис. 1. «мочь» как неправильно распознанная форма слова «моча»);

Б) своеобразно распознает запрос. Например, одним из параметров рассматриваемого запроса был родительный падеж, однако

фрагмент выдачи на рис. 1 демонстрирует, что коллокаты чаще всего не связаны с ключевым словом подобной связью (очевидно, что слово «духота» встречалось в тексте как однородный член предложения, а не коллокат в родительном падеже к слову «вонь»). Данный случай указывает на проблемы в распознавании синтаксических связей.

Данные примеры носят частный характер, но при анализе большего количества материала путем ручной проверки приведенных примеров, становится очевидно, что все затруднения можно условно разделить на следующие группы: синтаксические и лексические. К синтаксическим ошибкам отнесем неправильное понимание падежей, однородных членов и т.д. К лексическим — неправильную лемматизацию, проблемы в распознании сложных слов. Данная проблема чаще встречаются с прилагательными (на рис. 2. можно видеть, что автоматический разбор не воспринимает «удушливо-спертый» как одно прилагательное и учитывает только вторую часть).

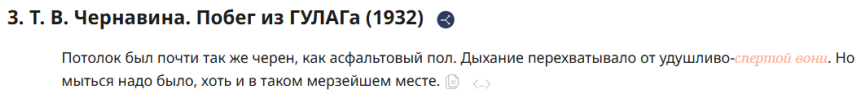


Рис. 2. Поиск прилагательных к слову «вонь»

Таким образом, уже на этом уровне заметны неточности, которые могут только увеличиться на большем объеме. Для исследователей важно понимать: являются ли эти ошибки чисто технической проблемой, которую можно решить путем настройки программного компонента, или принципиальной неспособностью технологий распознавать сложные синтаксические и тем более семантические конструкции?

В первом случае имеет смысл продолжить работу по улучшению инструментов корпуса, с целью доведения их до большей точности. Во втором случае становится очевидно, что квантитативные исследования с применением корпусов требуют осторожности и ограничений, т.к. автоматический синтаксический анализ текста не гарантирует абсолютной точности в распознавании зависимых слов и словосочетаний, вследствие чего выборка может содержать не все требуемые в запросе лексические единицы, либо некорректные данные.

Литература

Национальный корпус русского языка [Электронный ресурс]. Режим доступа: http://www.ruscorpora.ru/.

Павельева Т. Ю. Изучение коллокаций на основе лингвистических корпусов текстов // Вестник Тамбовского университета. Серия: Гуманитарные науки. 2016. С. 56–61.

Палийчук Д. А. Корпусные технологии в изучении коллокаций (на примере сервисов «antconc» и «sketchengine») // Studia Humanitatis. 2020. № 2.

Е.Н. Виноградова
(Москва, Россия)
МГУ имени М.В. Ломоносова
ekaterinavin@mail.ru

ДАННЫЕ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА ПРИ ОПИСАНИИ ФУНКЦИОНАЛЬНО-ГРАММАТИЧЕСКОГО ПОЛЯ ПРЕДЛОГА

В докладе продемонстрированы результаты применения данных Национального корпуса русского языка к описанию функционально-грамматического поля русского предлога; показано, что ранжирование предложных единиц по убыванию частотности позволяет стратифицировать данное поле от более грамматикализованных, более частотных единиц к менее грамматикализованным, окказиональным.

Ключевые слова: предлог, предложная единица, частотность, функционально-грамматическое поле предлога

Функционально-грамматическое поле (ФГП) русского предлога включает широкий круг средств предложного типа — как собственно предлогов, так и единиц, способных в определенных условиях выполнять функции предлога. Реестр предложных единиц однозначно не определен и, судя по всему, не может быть представлен в виде закрытого списка в силу постоянно идущих процессов грамматикализации различных типов сочетаний, пополняющих ФГП предлога.

Данные Национального корпуса русского языка (НКРЯ) могут быть использованы для объективации описания ФГП предлога и стратификации ПЕ от более грамматикализованных, ядерных к менее грамматикализованным, периферийным. Собранный в результате анализа различных словарей, грамматик и справочников реестр русских предложных единиц (подробнее [Виноградова 2017]) был разделен по структурным группам: первообразные предлоги и мотивированные ПЕ, среди которых противопоставляются отыменные, отадъективные и отглагольные ПЕ. Отдельные группы составляют компаративы, заимствованные ПЕ, а также наречные конкретизаторы и параметрические единицы. Внутри отыменных ПЕ были описаны как представленные предлогами в реестре РГ-80 модели (тв.п., р.п., в.п.; *без*, *в* + пр.п., *в* + в.п., *вне*, *за* + в.п., *за* + тв.п., *на* + в.п., *на* + пр.п.,

по + в.п, *по* + д.п., *под* + тв.п., *при, с* + р.п., *с* + тв.п., *с* + в.п., *через*), так и не нашедшие отражения в списке предлогов РГ-80 единицы.

Распределенные по указанным группам ПЕ были охарактеризованы по частотности, далее был определен суммарный ipm всех ПЕ в рамках каждой модели (табл. 1). Отметим, что суммарный ipm является более надежным показателем, чем число ПЕ, образованных по модели, так как есть группы, для которых выявлено достаточно много ПЕ, характеризующихся, однако, низкой частотностью.

Таблица 1. Структурные группы мотивированных предлогов, ранжированные по убыванию суммарной частотности, ipm

	модель	число ПЕ	суммарный ipm
1.	в+пр.п.	348	2001,36
2.	в+в.п.	371	1553,7
3.	по+д.п.	195	646,91
4.	отглаг.	60	539,66
5.	на+в.п.	95	411,33
6.	на+пр.п.	113	314,71
7.	отадъект.	58	247,42
8.	тв.п.	48	212,78
9.	с+р.п.	36	185,73
10.	с+тв.п.	30	180,34
11.	под+тв.п.	120	178,82
12.	нар.конкр.	37	160,13
13.	при	26	121,33
14.	за+в.п.	26	79,39
15.	не	33	67,80
16.	из	101	62,59
17.	до	13	54,02
18.	в.п.	8	51,81
19.	к	64	46.49
20.	за+тв.п.	40	49,35
21.	по+пр.п.	31	41,06
22.	с+в.п.	1	28,79
23.	от	10	27,5
24.	по+в.п.	23	25,76
25.	без	43	24,02

	модель	число ПЕ	суммарный ipm
26.	под+в.п.	81	23,19
27.	р.п.	11	20,69
28.	перед	9	14,71
29.	вне	40	7,76
30.	для	5	4,59
31.	заимств.	23	3,63
32.	через	7	2,64
33.	над	3	2,45
34.	о	2	2,07
35.	из-под	14	2,01
36.	сквозь	7	1,97
37.	у	4	1,21
38.	из-за	3	0,05
39.	компаративы	49	

Оказалось, что из наиболее частотных первообразных предлогов для «производства» вторичных предлогов не используется *про*. Единичные средства предложного типа образуются от предлогов *из-под, до, от, перед, сквозь, через, для, у, из-за, над, о, с*+в.п. и беспредложных форм р.п. и тв.п. Наиболее частотные предлоги *в, на, с, по* образуют максимальное число ПЕ, в то время как № 4 по частотности среди первообразных предлогов *к* мало используется для формирования вторичных ПЕ. Не самые частотные первообразные предлоги *под* и *при* «опережают» по потенциалу образования ПЕ более частотные *из, о, у, за, от, для, до* (таблица 1).

Наиболее продуктивной моделью является *в* + в.п. (371 ПЕ), однако по суммарной частотности «лидирует» *в* + пр.п. (ipm 2001,36). Очень частотны и многочисленны также ПЕ, образованные по разным моделям с помощью предлогов *на* и *по*. Думается, что это подтверждает мысль о значительной роли аналогии при образовании ПЕ, что способствует возникновению новых ПЕ в рамках «приоритетных моделей» и конкретных семантических групп.

Анализ частотности ПЕ, полученный по данным основного корпуса НКРЯ, позволяет выявить полевую устроенность ПЕ конкретных структурных групп русских предлогов и обосновать расположение ПЕ внутри функционально-грамматического поля предлога. Почти в

каждой структурной модели есть набор ядерных единиц с высокой частотностью и большая зона периферии, заполненная единицами со значительно более низкой частотностью, образованными по аналогии с ядерными, либо парадигматически связанными с ними: например, *по причине* — ipm 11,91, *по причинам* — ipm 0,31; *навстречу кому* — ipm 26,26, *навстречу к кому* — ipm 0,76; *с точки зрения* — ipm 13,52, *с позиции* — ipm 1,5.

Литература

Виноградова Е.Н. Проблемы лексикографического и грамматического описания предлогов в современном русском языке // Вопросы языкознания. 2017. № 5. С. 56–74.

РГ-80 — Русская грамматика / под ред. Н.Ю. Шведовой. М.: Наука 1980.

[1]Т.А. Гарипов, [2]А.В. Глазкова, [3]Я.Н. Губарькова,
[4]А.Д. Козеренко, [5]Д.А. Морозов
[1,5](Новосибирск, Россия)
[1,5]НГУ
[2](Тюмень, Россия)
[2]ТюмГУ
[3,4](Москва, Россия)
[3]Яндекс
[4]ИРЯ им. В. В. Виноградова РАН
[5]НП «НКРЯ»
[1]garipov154@yandex.ru, [2]a.v.glazkova@utmn.ru, [3]karmastina-ya@yandex-team.ru, [4]akozerenko@mail.ru, [5]morozowdm@gmail.com

УПРОЩЕНИЕ СЛОВАРНЫХ ТОЛКОВАНИЙ СЛОВ РУССКОГО ЯЗЫКА С ПОМОЩЬЮ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Непрерывное развитие русского языка требует поддержания актуальных толковых словарей. Развитие больших языковых моделей позволяет рассмотреть возможность автоматизации этой работы. В данной работе проведено сравнение различных подходов к генерации толкований с использованием современных языковых моделей и толковых словарей. Разработанный подход может быть использован для автоматизированного обновления и упрощения словарных статей с сохранением высокого качества определений, что повышает доступность и понимание информации для широких аудиторий.

Ключевые слова: генерация текста, большие языковые модели, обработка естественного языка, создание толковых словарей.

Постоянные изменения в социальной и культурной среде оказывают значительное влияние на различные аспекты человеческой деятельности, включая язык. Эти изменения особенно динамично проявляются на уровне лексики, где непрерывно возникают новые слова и заимствования из других языков, тогда как существующие лексические единицы подвержены семантической трансформации или постепенно выходят из употребления. В этом контексте толковые словари выполняют важную функцию, обеспечивая пользователей актуальными значениями слов, часто сопровождаемыми дополнительными лингвистическими пометами. Однако традиционные, вручную составляемые толковые словари ограничены в своей способности оперативно адаптироваться к изменениям в языке, что

связано с трудоемким характером их поддержки, требующей постоянного внимания со стороны лингвистов, а также значительных финансовых ресурсов.

В качестве альтернативы можно рассмотреть применение больших языковых моделей (LLM), которые обучены на обширных корпусах текстов и, теоретически, способны улавливать тонкости употребления лексем в разных контекстах и значениях [August et al. 2022, Malkin et al. 2021]. Более того, такие модели способны к генерации текстов в разнообразных стилях, включая упрощенные форматы, что делает возможным создание определений, доступных даже для детской аудитории. Такой подход к автоматизированному созданию толкований является новым для русскоязычных словарей и открывает возможности для дальнейшего исследования в области лексикографии и справочной информации.

Предыдущие эксперименты [Гарипов, Морозов 2024; Гарипов 2024] показали, что генерация толкований без опоры на первоисточник не позволяет добиться качественных и стабильных результатов. В связи с этим было решено перейти к задаче суммаризации и перефразирования набора толкований, а также материалы сайтов, расположенных на первых позициях в результатах поисковых систем по соответствующим запросам. Были протестированы следующие большие языковые модели: Vikhr-Nemo-12B-Instruct[1], Gemma-2-9b-it[2], Aya-expanse-8B[3], Qwen2.5-7B-Instruct[4], T-lite-instruct-0.1[5] и Llama-3.1[6] в различных конфигурациях.

В работе исследованы и протестированы различные подходы к построению взаимодействий между моделями, включая параллельную генерацию и сравнение моделью как судьей, а также раздельное использование моделей для создания и корректировки и дополнения определений. Для достижения максимальной полноты и грамотности толкований проводилась настройка промптов, что позволило повысить стабильность результатов и достичь более полной, логичной и структурированной генерации определений. Эксперименты показали, что модели способны производить определения с небольшим числом

[1] https://huggingface.co/Vikhrmodels/Vikhr-Nemo-12B-Instruct-R-21-09-24
[2] https://huggingface.co/google/gemma-2-9b-it
[3] https://huggingface.co/CohereForAI/aya-expanse-8b
[4] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct)
[5] https://huggingface.co/AnatoliiPotapov/T-lite-instruct-0.1
[6] https://openrouter.ai/nousresearch/hermes-3-llama-3.1-405b:free

грамматических ошибок, а также обеспечивать высокую точность и логичность структуры.

Разработанный подход демонстрирует потенциал для создания инструмента, способного автоматизировать процесс обновления словарей. Тем не менее, для полноценной оценки и повышения надежности предложенной методики требуется более масштабное тестирование. Такое тестирование планируется провести на платформе Национального корпуса русского языка [НКРЯ], что позволит более детально оценить применимость и качество предложенного решения. В дальнейшем данный подход может быть адаптирован для создания специализированных словарей и для задач анализа языковых изменений.

Литература

Гарипов Т. А. Генерация толкований с использованием больших языковых моделей // Всероссийская конференция «Путь в науку: прикладная математика, информатика и информационные технологии». Ярославль, 2024.

Гарипов Т. А., Морозов Д. А. Оценка качества толкований, сгенерированных с использованием LLM // Математическое и информационное моделирование. Материалы Всероссийской конференции молодых ученых. Тюмень, 2024.

Национальный корпус русского языка [Электронный ресурс]. URL: https://ruscorpora.ru/.

August, T., Reinecke, K., & Smith, N. A. Generating Scientific Definitions with Controllable Complexity // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022. С. 8298–8317. DOI: 10.18653/v1/2022.acl-long.569.

Malkin, N., Lanka, S., Goel, P., Rao, S., & Jojic, N. GPT Perdetry Test: Generating new meanings for new words // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021. С. 5542–5553.

А. С. Глаголева
(Москва, Россия)
Институт русского языка им. В. В. Виноградова РАН
glagoleva.anastasiia@mail.ru

ВОЗМОЖНОСТИ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА ПРИ ИССЛЕДОВАНИИ ЧАСТИЦ В РУССКОМ ЯЗЫКЕ

Русские частицы представляют довольно разнородную группу языковых единиц, изучение которых связано сразу с несколькими аспектами: во-первых, с определением закономерностей употребления частиц в стилистически и хронологически разных контекстах; во-вторых, с проблемой дифференциации частиц и других частей речи; в-третьих, с проблемой выделения категориального значения частиц как части речи. Современные корпусные методы, и, в частности, инструменты НКРЯ, значительно упрощают названные исследовательские задачи.

Ключевые слова: частицы, части речи, корпусная лингвистика, диахронические исследования

Согласно определению М. Г. Щур, частицы — это «одна из служебных частей речи, класс неизменяемых слов, выражающих многообразные отношения, реализующиеся в акте речи или тексте, а именно: отношение сообщаемого к участникам акта речи (к говорящему, слушающему), а также отношения между ними; отношение сообщаемого к действительности (в плане его реальности/ ирреальности, достоверности/недостоверности); отношение между высказываниями и их компонентами» [Щур 2021: 838].

Ввиду особой роли частиц в текстах разных типов, перед исследователем возникает задача выяснения особенностей их функционирования на материале заданного набора контекстов, составление которого упрощается благодаря инструментам, представленным в НКРЯ. Прежде всего, это параметры лексико-грамматического поиска: для отображения современных контекстов возможно указать искомую форму в поле «Лемма»; для исследования частиц (производных) в диахронии целесообразно задать поиск по словоформе, в особенности, если стоит задача определения хронологических рамок перехода полнозначной лексемы в частицу.

При совпадении полнозначной формы с частицей важную роль для их дифференциации выполняет морфологическая разметка корпуса и возможность поиска по критерию «Часть речи». Особенно это актуально для подкорпусов со снятой омонимией. Например, для частицы *хорошо* (при заданном наборе вариантов: лемма — *хорошо*, часть речи — *part*), случайная выдача примеров в корпусе со снятой омонимией (Основном) показывает достоверный результат — в 45 из 50 контекстов часть речи размечена правильно; в корпусе с неснятой омонимией (Мультимедийный) точность значительно ниже — при тех же параметрах поиска только 19 из 50 контекстов содержат частицу *хорошо*, а не омонимичную ей адвербиальную единицу.

Стоит, однако, сказать о том, что хотя возможности НКРЯ предоставляют исследователю эффективный инструмент поиска, при изучении частиц необходимо учитывать выходящую за пределы корпусной лингвистики проблему омонимии частиц другим частям речи. Так, по замечанию Д. В. Сичинавы, «поскольку служебные слова (предлоги, союзы и частицы) не изменяются, тем самым они по чисто морфологическому критерию слабо охарактеризованы и не могут быть отделены друг от друга и от наречий (если не относить сравнительную степень к наречию)» [Сичинава 2018: 23] Указанная проблема, помимо наречий, относится также к союзам и междометиям; для частиц, формирующихся в языке текущего момента, важно также отграничение их от глагольных форм. Эта проблема обусловлена, в том числе, отсутствием единого мнения о категориальном значении частиц — долгое время состав частиц определялся по остаточному принципу, и хотя в современной лингвистике решение этого вопроса предлагается посредством рассмотрения дискурсивных и прагматических функций частиц, разнообразие этих функций становится препятствием для выделения собственного значения частиц как части речи.

Для диахронных исследований частиц в НКРЯ иллюстративной является статистическая информация о результатах поиска и построение графиков. Например, при определении сравнительной частотности употребления частиц *ишь* и *вишь* график показывает, что пик употребления *вишь* приходится на середину XIX в., а затем частота снижается и более употребительной частицей становится *ишь*, что позволяет провести наблюдение над тенденциями развития обеих частиц.

Исследование особенностей употребления и значения конкретной частицы полезно также проводить в сравнении с данными других языков в параллельных подкорпусах. Так, например, обнаруживается, что частица *вишь* передается в других славянских языках с помощью языковых единиц со значением зрительного восприятия (*бач*, *виж*, *глей*), что близко к значению диахронной основы в русском языке (*видеть*); частица *именно* в иноязычных текстах передается средствами с другим значением и другим корнем (*точно*, *саме*).

Возможность поиска сочетаний слов позволяет изучить контекстуальные условия функционирования отдельных частиц. Например, выясняется, что частица *глянь* вплоть до конца XX века употребляется в сочетании с компонентом *-ка* (*-ко*), а с начала XXI века более употребительными становятся фразеологизированные сочетания (*куда ни глянь* и др.; подробнее об этом — [Глаголева 2024]).

Таким образом, инструменты, доступные в НКРЯ, могут быть полезны для широкого спектра лингвистических исследований, посвященных частицам.

Литература

Глаголева А. С. Глаголы со значением зрительного восприятия как основа для формирования частиц // Известия РАН. Серия литературы и языка. 2024. Т. 83. № 5. С. 137–149. (в печати)

Сичинава Д. В. Части речи // Материалы к корпусной грамматике русского языка. Выпуск III : Части речи и лексико-грамматические классы. СПб.: Нестор-История, 2018. С. 9–39.

Щур М. Г. Частицы // Русский язык: энциклопедия / Гл. ред. А. М. Молдован. 3-е изд., перераб. и доп. М.: АСТ-ПРЕСС ШКОЛА, 2020. С. 838–840.

[1]А. В. Глазкова, [2]Д. А. Морозов, [3]О. А. Митрофанова, [4]С. О. Савчук
[1](Тюмень, Россия)
ТюмГУ
[2](Новосибирск, Россия)
НГУ
[3](Санкт-Петербург, Россия)
СПбГУ
[4](Москва, Россия)
ИРЯ им. В.В. Виноградова РАН
[1]a.v.glazkova@utmn.ru, [2]morozowdm@gmail.com,
[3]o.mitrofanova@spbu.ru, [4]savsvetlana@mail.ru

ГЕНЕРАЦИЯ КЛЮЧЕВЫХ СЛОВ ДЛЯ ТЕКСТОВ РЕГИОНАЛЬНЫХ СМИ С ПОМОЩЬЮ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

В докладе представлен процесс разметки ключевых слов текстов региональной прессы Национального корпуса русского языка с помощью больших языковых моделей (LLM). Настройка модели осуществлялась на основе запросов с примерами текстов и их ключевых слов, что позволило добиться более точного, связного и полного подбора ключевых слов с меньшим количеством ошибок. Представленный подход планируется использовать для обновления разметки корпуса региональных средств массовой информации (СМИ).

Ключевые слова: генерация ключевых слов, Национальный корпус русского языка, газетный корпус, prompt-based learning, большая языковая модель.

Автоматическое выделение ключевых слов упрощает структурирование информации и обеспечивает быструю оценку содержания текстов. Ключевые слова могут использоваться для дальнейшего индексирования, рубрикации, реферирования и упрощения документов [Шереметьева, Осминин 2015; Ванюшкин, Гращенко 2018; Митрофанова, Гаврилик 2022]. В условиях большого потока информации выделение ключевых слов особенно важно для разметки текстов СМИ, поскольку оно облегчает поиск актуальных тем, улучшает навигацию по материалам и способствует проведению тематического анализа текстов.

В докладе представлен процесс генерации ключевых слов для текстов региональной прессы, включенных в Национальный корпус

русского языка [Савчук и др. 2024]. Формирование корпуса региональных СМИ началось в рамках международного проекта, реализуемого исследователями ИРЯ им. В. В. Виноградова РАН и Гродненского государственного университета им. Я. Купалы [Савчук 2015]. В настоящее время корпус насчитывает более 110 тыс. текстов (35 млн словоупотреблений). География корпуса охватывает все федеральные округа Российской Федерации и ряд стран ближнего зарубежья. Тексты имеют обширный набор метаатрибутов, включая тематику и тип текста, что позволяет организовать выбор подкорпуса. Поскольку категории текстов, сгруппированные по тематике или типу, представляют собой достаточно объемные подвыборки, для детального анализа узких тематических категорий также организован поиск текстов по ключевым словам.

По состоянию на октябрь 2024 года ключевые слова в текстах корпуса региональных СМИ размечены с помощью модели RuTermExtract (RTE)[1] и дополнительно обработаны с помощью ряда эвристик. В частности, были проведены нормализация словосочетаний с использованием набора правил и удаление коротких и однословных ключевых слов, являющихся личными именами. Ключевое слово может быть представлено униграммой или биграммой. Например, для текста о готовящейся июньской выставке кукол народов Приамурья список ключевых слов включает слова “*кукла*”, “*выставка*”, “*июнь*” и словосочетания “*уникальные экспонаты*”, “*традиционная кукла*”, “*творческая встреча*”. Разметка ключевых слов позволила более гибко описывать тематику текста и проводить ручной анализ ограниченной подвыборки, заданной набором ключевых слов. Использующая в настоящий момент модель разметки ключевых слов, основанная на RTE, имеет ряд недостатков, которые ограничивают возможности использования ключевых слов в корпусе региональных СМИ [Glazkova et al., 2024]. Так, модель часто генерирует общие слова, не описывающие тематику конкретного текста (“*день*”, “*участие*”), большое количество однокоренных слов, а также совершает ошибки, связанные с нормализацией словосочетаний (“*текстильные кукла*”, “*творческая*

[1] https://github.com/igor-shevchenko/rutermextract

мастерские”). Указанные недостатки потенциально преодолимы с помощью языковых моделей, предварительно обученных на больших текстовых корпусах и способных генерировать согласованные и содержательные тексты в ответ на поступающие в них запросы.

Для генерации ключевых слов была выбрана модель T-lite-instruct-0.1[1], которая относится к классу LLM, основанных на инструкциях (instruction-based). Настройка модели проводилась с помощью запроса (prompt) с уточнениями, что желаемое количество ключевых слов для текста составляет от пяти до десяти слов или выражений и ключевые слова требуется вывести в порядке уменьшения их значимости. Также запрос содержал несколько примеров текстов из корпуса региональных СМИ и соответствующих им списков ключевых слов. Ключевые слова для текстов-примеров были подобраны таким образом, чтобы они отражали основное тематическое содержание текста и его предметную область, то есть так, как это обычно делается при составлении аннотаций и ключевых слов к научным статьям. От модели требовалось вывести строку, представляющую собой список ключевых слов, разделенных запятыми. Эмпирическая оценка результатов показала, что ключевые слова, сгенерированные с помощью LLM, отличаются полнотой и связностью, а также содержат меньше грамматических ошибок по сравнению с результатами RTE.

В целом эксперимент по генерации ключевых слов с помощью LLM можно считать успешным. Благодаря способности эффективно обрабатывать и генерировать тексты на естественном языке, удалось обеспечить более точный подбор ключевых слов, которые отличаются полнотой, связностью и меньшим числом грамматических ошибок. В дальнейшем планируется использовать рассмотренный подход для обновления разметки ключевых слов в корпусе региональных СМИ.

Литература

Шереметьева С. О., Осминин П. Г. Методы и модели автоматического извлечения ключевых слов // Вестник ЮУрГУ. Серия: Лингвистика. 2015. Том 12. №1. С. 76–81.

[1] https://huggingface.co/AnatoliiPotapov/T-lite-instruct-0.1

Ванюшкин А. С., Гращенко Л. А. О разметке корпусов текстов ключевыми словами // Новые информационные технологии в автоматизированных системах. 2018. №21. С. 207–211.

Митрофанова О. А., Гаврилик Д. А. Эксперименты по автоматическому выделению ключевых выражений в стилистически разнородных корпусах русскоязычных текстов // Terra Linguistica. 2022. Том 13. №4. С. 22–40.

Савчук С. О., Архангельский Т. А., Бонч-Осмоловская А. А., Донина О. В., Кузнецова Ю. Н., Ляшевская О. Н., Орехов Б. В., Подрядчикова М. В. Национальный корпус русского языка 2.0: новые возможности и перспективы развития // Вопросы языкознания. 2024. №2. С. 7–34.

Савчук С. О. Корпус региональных газет России и зарубежья // Труды Института русского языка им. В.В. Виноградова. 2015. № 6. С. 163—193.

Glazkova A., Morozov D., Garipov T. Key Algorithms for Keyphrase Generation: Instruction-Based LLMs for Russian Scientific Keyphrases // arXiv 2410.18040. 2024.

[1]Д.А. Девяткин, [2]В.А. Салимовский, [3]Н.В. Чудова
[1,3](Москва, Россия)
ФИЦ «Информатика и управление» РАН
[2](Пермь, Россия)
Пермский государственный национальный исследовательский университет
[1]devyatkin@isa.ru, [2]salimovsky@rambler.ru, [3]nchudova@gmail.com

ПРИНЦИП СОЗДАНИЯ ОБУЧАЮЩЕЙ ВЫБОРКИ ДЛЯ БОЛЬШОЙ ЯЗЫКОВОЙ МОДЕЛИ

В докладе показано, что лингвистические признаки, фиксируемые во внутреннем состоянии большой языковой модели (БЯМ), точно воспроизводят особенности речевой системности обучающей выборки. Обосновывается мысль о том, что при решении классификационных задач, связанных с когнитивным моделированием, обучающую выборку целесообразно создавать из текстовых фрагментов, характеризующихся той разновидностью речевой системности, которая воплощает изучаемые когнитивные процессы.

Ключевые слова: большая языковая модель, обучающая выборка, речевая системность, признаковое пространство, гештальт.

В работах последних лет, интерпретирующих внутреннее состояние БЯМ [Zhu 2020, Pavlick 2022 и др.], показано, что анализ векторных представлений нейронной сети позволяет выявить отдельные лингвистические характеристики текстов обучающей выборки — принадлежность слов к определенной части речи, их синтаксические, семантические и риторические связи. Однако авторы этих работ не ставили перед собой задачу обнаружения целостной системы разноуровневых лингвистических признаков текста, поскольку не были понятны перспективы такого исследования.

Решая указанную задачу, мы исходили из предположения о том, что признаковое пространство БЯМ формируется в результате отображения речевой системности текстов обучающей выборки. Речевая системность понимается как взаимосвязь и взаимозависимость разноуровневых языковых средств в тексте по горизонтали и по вертикали на основе выполнения ими единого коммуникативного задания [Кожина 2020]. Она создается особенностями выбора, повторения, размещения, комбинирования и трансформирования языковых единиц. При этом разновидности речевой системности отличаются друг от друга «существенными различиями частот языковых

единиц и категорий… достаточными для их суммарного качественного опознавания…» [Головин 1971: 130].

Материалом исследования послужили научные тексты, являющиеся хорошим «экспериментальным полем» для проверки гипотез в области когнитивного моделирования. Рассматривались два речевых жанра — «Описание нового для науки явления» и «Экспликация научного понятия». Первый из них воплощает эмпирическую, а второй — теоретическую познавательную деятельность. В состав каждой из двух обучающих выборок размером свыше 7 тыс. словоупотреблений включались только те фрагменты текстов, которые представляют познавательные действия, точно соответствующие номинации речевого жанра. Цель экспериментов состояла в сравнении лингвистических признаков, устанавливаемых в результате анализа векторных представлений БЯМ (глубокой нейронной сети с архитектурой «Трансформер»), с признаками речевой системности обучающих выборок.

Программный метод состоял в построении схем предложений исследуемого текста на основе внутренних векторных представлений нейронной сети и в установлении фрагментов этих схем, связанных с целевыми результатами анализа текста. Схемы включали словоупотребления и синтаксические связи между ними.

Анализ показал, что состав словоформ и частота их употребления, отображенные во внутреннем состоянии БЯМ, полностью совпадает с составом и частотностью словоформ обучающих выборок. Иными словами, векторные представления точно фиксируют состав и частоту употребления слов в тех или иных их морфологических формах и синтаксические позиции словоформ в составе предложений, т.е. речевую системность этих выборок на лексическом, морфологическом и синтаксическом уровнях.

Механизм работы лингвистического модуля БЯМ характеризуется нами так: предварительное обучение нейронной сети на больших (в миллиарды словоупотреблений) корпусах текстов призвано запечатлеть в ней систему языка — его единиц и общих принципов их использования, а дообучение на корректно составленных обучающих выборках — перестройку системы языка в систему речи. (О соотношении языковой и речевой системности см.: [Кожина 2020: 196-198].)

Поскольку в наших экспериментах речевая системность обучающих выборок являлась именно такой организацией речи, которая воплощает изучаемые когнитивные процессы (фиксацию признаков нового явления и экспликацию понятия), качество распознавания

нейронной сетью каждого из рассматриваемых речевых жанров оказалось ожидаемо высоким. Так, в то время как показатели F_1-меры предварительно обученной БЯМ (модели, отражающей только систему языка) были соответственно 0.69 и 0.47, показатели дообученной БЯМ (модели, запечатлевшей речевую системность) — в обоих случаях 0.99.

Проведенные эксперименты позволяют сформулировать принцип создания обучающей выборки при решении БЯМ классификационных задач, связанных с когнитивным моделированием: **выборку целесообразно составлять из текстовых фрагментов, речевая системность которых воплощает изучаемые ментальные процессы**.

Что представляет собой признаковое пространство (классификационное основание) во внутреннем состоянии БЯМ? Результаты экспериментов позволяют предположить, что им является образ «воспроизводимых частот языковых единиц и категорий» (Б.Н. Головин), узнаваемый не только человеком, но и машиной, — своего рода гештальт. Действительно, устранение из обучающей выборки отдельных лингвистических характеристик текстовых фрагментов (например, в тексте именного типа — всех глаголов, или всех наречий, или глаголов и наречий вместе) не приводит к изменению показателя F_1-меры. В связи с этим актуальной становится задача экспериментального исследования речевой системности для определения границ гештальта при работе с БЯМ, а также для установления закономерностей ее формирования в нейросетевом обучении.

Литература

Головин Б.Н. Язык и статистика. М.: Просвещение, 1971. 190 с.

Кожина М.Н. Речеведение. Теория функциональной стилистики: избранные труды. М.: Флинта: Наука, 2020. 624 с.

Pavlick E. Semantic structure in deep learning //Annual Review of Linguistics. 2022. Vol. 8. P. 447–471.

Zhu Z. et al. Examining the rhetorical capacities of neural language models // Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. 2020. P. 16–32.

Е. В. Дзюба, О. А. Щербак
(Санкт-Петербург, Россия)
Санкт-Петербургский политехнический университет Петра Великого

«СЛУЖИТЬ БЫ РАД...»: НКРЯ — ЯЗЫКУ ДЛЯ СПЕЦИАЛЬНЫХ ЦЕЛЕЙ

В докладе освещается работа над проектом, реализуемым коллективом авторов — преподавателей русского языка (в том числе РКИ) Высшей школы международных отношений Гуманитарного института Санкт-Петербургского политехнического университета Петра Великого. Целью проекта является составление учебного словаря языка для специальных целей, рабочее название: «Язык политики и международных отношений. Учебный словарь». Спецификой словаря является то, что значительная часть материала для словарной статьи (близкие по значению слова, сочетаемость, примеры контекстов, некоторые грамматические особенности слов) черпается из Национального корпуса русского языка (основной и газетный корпуса) посредством инструмента «Портрет слова».

Ключевые слова: Национальный корпус русского языка, учебный словарь, словарь языка для специальных целей, проект учебного словаря, язык политики и международных отношений.

Национальный корпус русского языка (НКРЯ) — «очень эффективный и полезный инструмент, которым могут пользоваться далеко не только узкие специалисты»; он являет собой «собрание текстов на данном языке, представленное в электронной форме и снабженное … встроенным ... научным аппаратом», называемым «разметкой» [Плунгян 2005: 6]. В. А. Плунгян отмечает весьма широкий спектр возможностей использования корпуса в разных областях: от многообразных и разноаспектных лингвистических и филологических исследований — до технологических (по информатике и программированию), лингводидактических (особенно при усвоении и преподавании русского языка как иностранного), редакторско-журналистских и иных практик [см. подробнее: Плунгян 2005]. Развитие корпусных исследований ученый считает наиболее перспективным направлением в лингвистике XXI века, ее «точкой роста» [Плунгян 2018]. При поисковом запросе *НКРЯ* в электронной библиотеке elibrary.ru находим 87637 научных публикаций: таким образом, «обращение к Корпусу стало почти самоочевидным

компонентом практически любого исследования русского языка» [Плунгян 2015: 11]. Уместно использование корпусных данных и в лексикографической практике [Плунгян, Сичинава 2005].

Коллектив авторов, преподавателей русского языка (в том числе русского как иностранного) Высшей школы международных отношений Гуманитарного института Санкт-Петербургского политехнического университета Петра Великого (Е. В. Дзюба, С. А. Губарева, И. А. Краснова, Е. Н. Туана, О. А. Щербак), реализует проект, направленный на составление учебного словаря языка для специальных целей (ЯСЦ) с опорой на данные НКРЯ (рабочее название: Язык политики и международных отношений. Учебный словарь). Следует подчеркнуть, что сама идея составления словарей ЯСЦ по данным Национального корпуса русского языка может быть нереализуемой, поскольку в базу текстовых данных НКРЯ входят только учебно-научные, но не собственно научные тексты. Очевидно, что единицы языка далеко не каждой специальности (особенно терминологические) попадают в корпус, однако для сферы политологии и международных отношений ситуация складывается благоприятно: НКРЯ фиксирует социально-политическую и даже терминологическую лексику этой области знания преимущественно в газетном корпусе. Встречаются, однако, некотрые термины, которые по понятным причинам не отражаются в НКРЯ, ср.: *трипартизм*, «система трехстороннего представительства (государства, работников и работодателей) в процессе регулирования социально-трудовых отношений», *брачная дипломатия* и нек. др.; в таком случае данные для словарной статьи черпаются из иных источников. Но это касается не всех терминов: так, в НКРЯ представлены, например, такие единицы, как *легитимизм, колониализм, транспарентность* и мн. под.

Данный учебный словарь содержит социально-политическую и специальную лексику, актуальную для профессиональной подготовки студентов по направлению «Международные отношения и зарубежное регионоведение». Спецификой словаря является то, что значительная часть материала для словарной статьи (близкие по значению слова, сочетаемость, примеры контекстов, некоторые грамматические особенности слов) черпается из НКРЯ (основной и газетный корпуса) посредством инструмента «Портрет слова», а именно из следующих его виджетов — осовного корпуса: *Скетчи* (коллокации), *Формы*

слова, *Однокоренные слова*, *Похожие слова*; газетного корпуса (Центральные СМИ): *Скетчи*, *Похожие слова*, *Примеры*.

Словарная статья, таким образом, включает следующую информацию о слове: сведения об ударении и при необходимости произношении (ср.: *блоки́ровать*, *аг[рэ]ссия, а[ф'е]ра,*); грамматических характеристиках (особенно — морфологических особенностях, ср.: *колониализм*, только формы ед.ч.) и единицах словообразовательного гнезда (ср.: к слову *конфликт*: *конфликтовать*, *конфликтный*, *бесконфликтный*); значении, актуальном для сферы политологии, дипломатии и международных отношений (ср.: *колония* — страна, захваченная другим государством и лишённая политической и экономической самостоятельности); типичной сочетаемости (ср. коллокаты слова *конфликт*: *вооруженный*, *межнациональный*, *локальный*; *конфликт возникает* / *начинается* / *продолжается* / *обостряется* / *назревает*; *урегулировать* / *уладить* / *спровоцировать* / *разрешать конфликт* / *избегать конфликта*; *обернуться* / *сопровождаться* / *закончиться конфликтом*; *вмешиваться* / *вступать* / *вылиться* / *втянуть* / *перерасти в конфликт*; *участвовать в конфликте*, *привести к конфликту*); синонимах и антонимах к слову (ср. к слову *конфликт*: *столкновение*, *противостояние*, *разногласие*, *конфронтация*, *раздор*, *спор*; *мир*, *согласие*, *единение*, *примирение*); употреблении в контексте на уровне предложения (ср. иллюстрацию к слову *конфликт*: *На пике карьеры он ... занимался урегулированием межнациональных конфликтов*).

В заключение подчеркнем, что для создания учебного словаря специального языка сферы политологии и международных отношений материалы и инструменты НКРЯ весьма полезны. Не каждый студент (да и преподаватель) будет обращаться к ресурсам НКРЯ при знакомстве с тем или иным словом ЯСЦ. Подготовленный с привлечением данных НКРЯ учебный словарь поможет быстро и полноценно освоить основные «портретные» характеристики слова в учебной практике.

Литература

Плунгян В. А. Зачем нужен национальный корпус русского языка? Неформальное введение // Национальный корпус русского языка: 2003—2005. Результаты и перспективы. М., 2005. С. 6–20.

Плунгян В. А. Лингвистика в XXI веке: проблемы, перспективы, точки роста // Слово.ру: балтийский акцент. 2018. Т. 9. № 1. С. 7–12.

Плунгян В. А. Предисловие // Труды института русского языка им. В.В. Виноградова. 2015. № 6. С. 11–20.

Плунгян В. А., Сичинава Д. В. Национальный корпус русского языка как инструмент лексикографа // Слово и словарь. Гродно, 2005. С. 197–202.

Сичинава Д. В. Национальный корпус русского языка: очерк предыстории // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005.

Д.О. Добровольский
(Москва, Россия)
ИРЯ им. В.В. Виноградова РАН
dm-dbrv@yandex.ru

ПАРАЛЛЕЛЬНЫЙ КОРПУС И ДВУЯЗЫЧНЫЙ СЛОВАРЬ

В центре внимания исследования стоит вопрос, каким образом реальные употребления лексических единиц в параллельном корпусе во всем их многообразии могут быть сопоставлены с релевантными элементами структуры словарной статьи двуязычного словаря. Этот вопрос рассматривается в докладе на примере проекта по созданию ресурса, интегрирующего материал двуязычного — немецко-русского — словаря и данные параллельного немецко-русского корпуса, входящего в состав Национального корпуса русского языка[1].

Центральной оказывается при этом проблема многозначности лексических единиц. Если слова и фраземы, обладающие только одним значением, позволяют постулировать связь между словарной статьей и корпусными примерами на уровне леммы, то все многозначные единицы требуют соотнесения корпусных данных со словарной информацией на уровне каждого конкретного значения данной лексической единицы.

Исследование, направленное на выработку принципов соединения словарной и корпусной информации, осуществляется на двух фрагментах лексической системы. Это, с одной стороны, немецкие модальные глаголы с их русскими соответствиями, а с другой — устойчивые конструкции (фраземы) разной степени идиоматичности. Эти два фрагмента выбраны не случайно.

Модальные глаголы отличаются необыкновенно развитой полисемией, так что интеграция сведений из двуязычного словаря по каждому из них с обширным корпусным материалом представляет существенные трудности. Каждое конкретное употребления модального глагола в корпусе должно быть соотнесено с одним из его значений, представленном в словаре.

[1] Данный проект выполняется в настоящий момент в Федеральном исследовательском центре «Информатика и управление» Российской академии наук за счет гранта Российского научного фонда № 24-18-00155.

Часто это дополнительно затрудняется обилием русских переводных эквивалентов, выбор которых не всегда мотивирован собственно семантическими параметрами.

Сложности описания устойчивых конструкций мотивированы не столько количеством и разнообразием значений каждой фраземы, сколько принципиальной непредсказуемостью их синтаксического поведения. Само включение фразем в электронный словарь с обеспечением их поиска по различным параметрам — задача, к решению которой компьютерная лексикография начинает подходить только в самые последние годы.

Выбор указанных фрагментов лексикона (модальных глаголов и устойчивых конструкций) оправдан, таким образом, наличием у каждого из этих фрагментов своих особых характеристик. Можно предполагать, что если специфические трудности, связанные с обработкой каждого из этих фрагментов будут успешно решены, это будет означать, что проектируемая система способна справиться с любыми потенциальными трудностями, возникающими при попытке осуществить настройку словаря на корпус, а корпуса на словарь.

М. В. Дудорова
(Екатеринбург, Россия)
Уральский федеральный университет
им. первого Президента России Б.Н. Ельцина
primrose81@yandex.ru

О. И. Северская
(Москва, Россия)
Институт русского языка им. В. В. Виноградова РАН
oseverskaya@mail.ru

КОРПУСНЫЕ ДАННЫЕ КАК КЛЮЧ К ФИКСАЦИИ СОВРЕМЕННЫХ ГРАММАТИЧЕСКИХ ПРОЦЕССОВ (на материале новых СКС на базе сочетаний с предлогами *в/на/по*)

В предлагаемом вниманию исследовании рассматривается процесс грамматикализации предложно-падежных сочетаний с предлогами *в, на* и *по* в русском языке. Анализируются структурные, лексико-семантические, стилистические характеристики и парадигматические отношения этих сочетаний. Особое внимание уделяется их функциональной нагрузке и использованию в различных синтаксических позициях. Представлена классификация рассмотренных предложно-падежных форм по семантическим группам и подгруппам, а также их словарная фиксация и отражение в Национальном корпусе русского языка (НКРЯ).

Ключевые слова: слова категории состояния; грамматикализация; предложно-падежные сочетания; корпусные исследования.

Исследование, о котором пойдет речь, представляет собой определенный вклад в развитие «корпусной грамматики», поскольку имеет целью изучение чрезвычайно распространившихся в последнее десятилетие предложно-падежных сочетаний с предлогами *в, на* и *по* в процессе их грамматикализации, которая проявляется, с одной стороны, в застывании формы, с другой стороны, в приобретении ими статуса СКС.

Объектом исследования, выполненного на материале НКРЯ в максимальном объеме представляющих синхронный срез языковой системы подкорпусов, стали 29 предложно-падежных сочетаний типа *на стиле, по кайфу, в моменте* и под., предметом — лексико-

семантические, структурные и функциональные аспекты употребления этих сочетаний.

1. Структурный аспект.

Рассматриваемые сочетания образуются по моделям: (*в*) N_4, (*на*) N_6, (*по*) N_3[1].

2. Лексико-семантический аспект.

2.1. Рассмотренные единицы представляют различные семантические группы (классификация приводится в соответствии с синопсисом, разработанным на основе комплекса идеографических словарей, подготовленных учеными Уральской семантической школы [Бабенко 2015]):

— наибольшее количество единиц представляют семантику денотативно-идеографической группы (далее — ДИГ) «Эмоции»: подгруппы «Общие понятия» (*на эмоциях*), «Беспокойство» (*на панике, в панике*), «Растерянность» (*в шоке*), «Удовольствие» (*в кайфе, по кайфу*), «Спокойствие» (*на чиле, на расслабоне, на релаксе, в принятии, в потоке, на лайте*[1]), «Радость» (*на позитиве, на драйве*), «Смелость» (*на кураже*), «Неприязнь» (*на негативе*), «Воодушевление» (*в ресурсе*);

— ряд единиц — семантику ДИГ «Интеллект»: подгруппы «Познание» (*в курсе*), «Память» (*на опыте*), «Восприятие» (*в адеквате*[1]*, в неадеквате*[1]), «Понимание» (*на волне*);

---

[1] Прототипические формы нестандартных предложно-падежных сочетаний имеют словарную фиксацию. Так, в «Универсальном словаре русского языка» под ред. В.В. Морковкина фиксируется выражение *в форме,* характеризующее «хорошее состояние, позволяющее человеку полностью проявить свои силы, умение, способности», с вариантами *быть в форме / не в форме* [Морковкин]. В «Словаре русского арго» В.С. Елистратова есть выражение *на стрёме*, означающее «состояние тревоги, беспокойства, бдительной готовности к чему-л.» [Елистратов]. А в «Словаре русского языка» в 4-х томах под ред. А.П. Евгеньевой зафиксировано «**на мази́** *что* (*прост.*) — в благоприятном положении, состоянии» [Евгеньева].

— ДИГ «Социальные качества и поведение человека», подгруппа «Человек и его поведение по отношению к нормам этикета» (*в адеквате*[2]*, в неадеквате*[2]);

— ДИГ «Оценка», подгруппа «Прагматическая, утилитарная оценка» (*на изи, на лайте*[2]);

— ДИГ «Одежда», подгруппа «Человек, одетый/неодетый каким-либо образом» (*на стиле*);

— ДИГ «Спорт», подгруппа «Общие понятия, связанные со спортом» (*на спорте*[1]),

— ДИГ «Человек как живой организм», подгруппа «Особенности телосложения человека» (*на спорте*[2]);

— ДИГ «Время», подгруппы «Время относительно момента речи» (*в моменте*) и «Всегда/никогда» (*по жизни*);

— ДИГ «Универсальные смыслы и отношения», подгруппы «Развитие каких-либо процессов, явлений, событий» (*в процессе*) и «Интенсивность» (*на минималках, на максималках*).

2.2. Стилистические характеристики. Рассматриваемые единицы фиксируются преимущественно в Устном корпусе и в корпусе «Социальные сети», что дает возможность введения помет *разг.* и *нейтр*.

2.3. Проявления парадигматических отношений

— синонимия (формы типа *в ресурсе, в потоке, в принятии* часто используются вместе, превращаясь в контекстные синонимы).

— антонимия обусловлена антонимией производящих существительных и образует пары: *в адеквате — в неадеквате; на позитиве — на негативе; на минималках — на максималках*.

3. Функциональный аспект.

Рассмотренные предложно-падежные формы употребляются в следующих синтаксических позициях (часто с глаголами бытия и состояния):

— в качестве присвязочной части составного сказуемого с замещением позиции Adj_f (*Мы всегда* ***на волне****; Я сегодня* ***на позитиве****; Вот же человеку всё* ***по кайфу***),

— в присловной позиции в качестве субститута в модели V + Adv (*решить* ***на панике***; *ответить* ***в моменте****; танцевать* ***на расслабоне****; сделать* ***на изи***),

— в позиции детерминанта (*Было очень весело, хотя* ***в моменте*** *нам так не казалось*).

На уровне словосочетания зарегистрированы: а) определительные (*он был* ***на волне*** = популярный; *пришел весь* ***на стиле*** = модный) и б) обстоятельственные (*я вообще* ***по жизни*** *пишу с ошибками*) синтаксические отношения.

Полученные результаты свидетельствуют о развитии рассмотренных сочетаний с предлогами *на, в* и *по*, их грамматикализации по типу СКС и их интеграции в языковую систему.

Литература

Бабенко Л. Г. Синопсис (свод) идеографической классификации русской лексики (общая глобальная структура словаря) // Универсальный идеографический словарь русского языка: проспект / под общ. ред. Л. Г. Бабенко. М.; Екатеринбург: Кабинетный ученый, 2015. С. 22–42.

Евгеньева А.П. (ред.). Словарь русского языка: В 4-х т. 4-е изд., стер. М.: Рус. яз.; Полиграфресурсы, 1999 [Электронный ресурс]. URL: Фундаментальная электронная библиотека (дата обращения 04.11.2024).

Елистратов В.С. Словарь русского арго [Электронный ресурс]. URL: slovari.ru (дата обращения 04.11.2024).

Морковкин В.В., Богачева Г.Ф., Луцкая Н.М. Большой универсальный словарь русского языка. URL: https://gramota.ru/biblioteka/slovari/bolshoj-universalnyj-slovar-russkogo-yazyka (дата обращения 04.11.2024).

А. А. Евдокимова
(Москва, Россия)
Институт языкознания РАН
arochka@gmail.com

ПРОБЛЕМЫ ВЕРИФИКАЦИИ ДАННЫХ В КОРПУСАХ ДРЕВНИХ ТЕКСТОВ

В докладе на примере надписей, папирусов, рукописей etc., вошедших в корпус акцентуированных византийских текстов BGAT, будут показаны разные подходы к сбору корпуса и верификации данных в нем. Особое внимание будет уделено анализу спорных случаев и проблеме подготовки фотографий и прорисовок письменных источников к разметке и дальнейшему тегированию для создания верификационной «3D модели» памятника.

Ключевые слова: корпус BGAT, древние тексты, эпиграфика, корпусная лингвистика, древнегреческий язык

В процессе сбора корпуса акцентуированных византийских текстов BGAT [Евдокимова 2023] мы столкнулись с тем, что многие исследователи издавали надписи и другие источники, не приводя их фотографий. Некоторые из авторов публикаций использовали помимо унифицированного текста и комментариев, дипломатическое издание. Другие, столкнувшись с невозможностью в печатном виде передать все особенности памятника, выполняли его прорисовки. Однако очень много публикаций византийских и античных греческих текстов (надписей, рукописей, папирусов, печатей, граффити, дипинти и т.д.), которые не были переизданы с дополнениями более поздних исследователей, были изданы с той или иной степенью унификации. При этом унифицировались не только знаки акцентуации в тех текстах, которые их содержали, но и ставились там, где их не было. Кроме того, некоторые издатели XIX и нач. XX веков предпочитали исправлять орфографические ошибки, допущенные в оригинале, или приводить текст в более «читабельный вид» согласно моде того времени.

Помимо унификации текста и отсутствия фотографий в изданиях другой проблемой при верификации данных оказывается разница подходов к восстановлению неполных или частично утраченных памятников. Кто-то идет методом аналогий и анализа формул в случае отсутствия достойных упоминания вариантов, приводит текст, как есть, указав размер лакун или передав максимально близко испорченное место. Кто-то предпочитает высказывать некоторую

теорию, которая подходит, по его мнению, по смыслу, предлагает восстановление в рамках ее, не воспроизводя в точности испорченное место или не указав размер утраченного. Сейчас, в эпоху общемировой цифровизации хранящихся в музеях фондов, все случаи восстановления под идею становятся все более уязвимыми, так как появляются недоступные ранее исследователям аналогии.

Не у всех исследователей, часто в силу не зависящих от них причин, есть возможность вернуться к уже изданным текстам и перепроверить собственное прочтение и/или прорисовку при изменении состояния памятника, например, после его реставрации, или с появлением более четкой фотофиксации, или новых методов фотофиксации, которым является разработанный сейчас способ создания 3D моделей [Свойский и др. 2018]. Как работает этот метод на практике, можно увидеть на примере самаритянской надписи: https://rssda.su/projects/ep-sam/ (дата обращения: 10.11.2024) или уже с детальным анализом на материале византийской надписи [Евдокимова и др. 2023].

Такая ситуация с изданиями памятников требует при создании корпуса не только поиска фотографий, снятых нашими современниками, в случае если сохранность памятника позволяет, но и использования всех возможных прорисовок и фотографий, хранящихся в музейных и библиотечных архивах и опубликованных в изданиях. Конечно, идеальным вариантом верификации результата может быть как раз выполнение 3D модели источника, если памятник сохранился и можно получить доступ для его 3D моделирования. Однако многочисленные случаи утраты памятников или их недоступности приводит к необходимости описывать не только текст по всем доступным изданиям, создавая критическое издание внутри корпуса, но и сопоставлять доступные фотографии и прорисовки прошлых лет. Т.е. в каком-то смысле составлять верификационную «3D модель» источников, перепроверяя все обстоятельства публикации, в том числе сопоставляя друг с другом памятники из одного издания по степени их достоверности. Как показал опыт анализа изданий греческих надписей Грузии, выполненных Т. Каухчишвили в 1951 и 1999-2002 годах, в каких-то случаях при создании прорисовки исследователь в первом издании выбирала путь унификации, не только знаков акцентуации, приводя их к византийской системе акцентуации, но и орфографии, трактуя в местах плохой сохранности в пользу общепринятой орфографии. Однако, ее более позднее издание, выполненное уже через 48 лет, дает иную картину для ряда памятников, и сравнение их с

небольшим числом доступных фотографий показало, что для многих надписей оно является более точным [см. Евдокимова 2024]. В корпус эти надписи после подобного сопоставления входят не только в виде цитат из издания, но и всех доступных прорисовок и фотографий, каждая из которых содержит теги, маркирующие все значимые для греческой акцентуации и орфографии факторы, а также сопутствующих исследовательских комментариев детально описывающих их лингвистические и палеографические особенности. Принципы тегирования подробно описаны нами в статье, посвященной презентации корпуса [Евдокимова 2023], и были апробированы на разных типах памятников, включая папирусы, надписи на металле, мозаики, рукописи, печати, надписи на фресках, граффити и т.п. В докладе будут показаны их верификационные «3D модели» и разные подходы к их составлению.

Литература

Евдокимова А. А. Корпус византийских письменных памятников и методы его разметки // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог» (2023), серия 22, издательство МИИ (Москва), том 22. С. 1071-1081. DOI: 10.28995/2075-7182-2023-22-1071-1081.

Евдокимова А.А. Гелати, греческие надписи на фресках и мозаиках и их акцентуация. // ByzantinoCaucasica. Выпуск 4: сб. ст. / Отв. ред. В. Н. Чхаидзе; Ин-т. востоковедения РАН — М.: ФГБУН ИВ РАН, 2024. С. (в печати)

Евдокимова А.А., Мастыкова А.В., Свойский Ю.М. Христианское надгробие с византийского памятника Горзувиты (южный берег Крыма) // Краткие сообщения Института археологии. вып. 271, 2023. С. 184-198. DOI: 10.25681/IARAS.0130-2620.271.184-198

Свойский Ю.М., Романенко Е.В., Миклашевич Е.А. Опыт создания цифровых образов эстампажей енисейских петроглифов методом трехмерного моделирования // Camera praehistorica. 2018. № 1 (1). С. 106–116. DOI: 10.33291/26583828.2018-(1)-6.

[1]А. Г. Жукова, [2]О. И. Северская
(Москва, Россия)
[1]Государственный институт русского языка им. А. С. Пушкина
[2]Институт русского языка им. В. В. Виноградова РАН
[1]arinazhukova2013@gmail.com, [2]oseverskaya@mail.ru

ПИСЬМА «ДЕЛЬНЫЕ» И «ДЕЛОВЫЕ» В ЭПИСТОЛЯРНОМ НАСЛЕДИИ А.С. ПУШКИНА И РУССКОЙ ЛИНГВОКУЛЬТУРЕ

В исследовании рассматривается семантика и употребление эпитетов «дельный» и «деловой» в письмах А.С. Пушкина в проекции на лексикографическую фиксацию до- и послепушкинской поры и современности. Анализируется рефлексия поэта относительно «дельности» и «вздора» в частной переписке. Результатом исследования становится корпусный портрет эпитетов «дельный» и «деловой», отражающий их особое место в эпистолярном прагматиконе.

Ключевые слова: корпусное исследование, портрет слова, эпистолярный прагматикон, эпистолярный кодекс А.С. Пушкина, эпитеты «дельный» и «деловой», семантическое развитие.

Эпистолярное наследство А.С. Пушкина занимает особое место в истории переписки в России: поэт лишь отчасти следует нормам своего времени и предлагает новые, которые впоследствии будут кодифицированы. Особый интерес представляют пушкинские метавысказывания о правилах переписки, его ожиданиях от адресата, о «хороших» и «плохих» письмах, а также характеризующие переписку эпитеты.

Репертуар эпитетов к слову *письмо*, используемых в пушкинском эпистолярии, довольно богат: нами из текстов, вошедших в эпистолярный подкорпус Основного корпуса НКРЯ и в корпус «Русская классика», выделено около 30 прилагательных, часть которых повторяется неоднократно, иные же встретились в единичном употреблении.

Обращают на себя внимание прилагательные с обще- и частнооценочными значениями [Арутюнова 1988: 92-96]. Первые однотипны: *прекрасное, милое / премилое / премиленькое письмо*. Вторые разнообразны, они транслируют этические, нормативные оценки письма (*благоразумное, порядочное, неприличное*); а также утилитарные

(*умное / глупое, бестолковое*), выражающие коммуникативную установку адресанта (*откровенное, решительное*) и эмоциональное восприятие читателя эпистолярного текста (*уморительное, смешное, вдохновительное, нежное, трогательное, сопливое, полу-кислое, премеланхоличное, грустное, горькое, мрачное* и др.) (здесь и далее сохраняется авторское написание). Еще одна группа прилагательных характеризует объем писем и наличие в них деталей и подробностей: *короткое, коротенькое* («плохое»), *длинное, подробное* («хорошее»). Довольно много «форматных» определений писем — по жанру (*обещательные, извинительные*), по коду (*официальные, партикулярные, дружеские*). К «форматным» эпистолярным эпитетам можно отнести и паронимическую пару *дельный — деловой*. Развитие семантики этих прилагательных в языке, в том числе в языке А.С. Пушкина, представляет особый интерес.

Прилагательное *дельный* одно из высокочастотных слов пушкинского идиолекта: оно использовано в текстах А.С. Пушкина 37 раз, из них 24 раза — в письмах. При этом употребляется не всегда в значениях, характерных для пушкинской поры.

В «Словаре Академии Российской» (1790) *дельный* имеет значения ‘работный, определенный для трудов’, ‘годный для дела’, ‘основательный’. В «Толковом словаре живого великорусского языка» В.И. Даля (1863) *дельный* значит ‘пригодный в дело, годный к чему-л.’ В пушкинских же письмах *дельный* употребляется в значениях ‘обстоятельный, основательный, содержательный, толковый’: *Очень обрадовался я, получив от тебя письмо (**дельное** по твоему обычаю). Постараюсь отвечать по пунктам и обстоятельно…* [А.С. Пушкин. Письмо П.А. Плетневу, около 11.10.1835]. Налицо развитие семантики, приближающее нас к современному значению, ср. в «Большом универсальном словаре русского языка» под ред. В.В. Морковкина: *дельный* ‘такой, к-рый отличается разумностью, существенностью, имеет практическое значение’, в «Большом толковом словаре русского языка» под ред. С.А. Кузнецова: *дельный* ‘касающийся существа, практически полезный, толковый’. Необходимо отметить, что 70% употреблений А.С. Пушкиным прилагательного *дельный* относится к речевым произведениям — как *дельные* оцениваются *записки, письмо, статья, замечания, критика, возражение* и т.д.

Среди найденных примеров обращает на себя внимание «когнитивно-стилистическая» рефлексия, сопоставляющая *дельность* и *вздор*: *Главная ошибка наша была в том, что мы хотели быть слишком* ***дельными****; стихотворная часть у нас славная; проза м. б. еще лучше, но вот беда: в ней слишком мало* ***вздору*** [А.С. Пушкин. Письмо М.П. Погодину, 31.08. 1827]. Подобное разграничение находит неожиданные переклички в переписке с близкими людьми, женой и братом: первую хвалит за дельное, длинное письмо, призывая при этом писать «часто и *о* всяком *вздоре*, до тебя косающимся», другому адресует отповедь: «не благодарю тебя за письмо твое, потому что ты мне *дельного* ничего не говоришь — я называю *дельным все, что касается до тебя*», обещая в свою очередь «отвечать *со всевозможной болтливостью». Дельность* для него — это факты в мельчайших подробностях, *вздор* — эмоции и «воздух», помогающий адресату комфортно войти в ритм письма и суть дела.

Если в «Словаре Академии Российской» и у В.И. Даля *деловой* — это прежде всего ‘искусный в приказных, письменных делах’ и ‘работающий по найму’ (о человеке) и ‘к исправлению работы относящийся’ (в других случаях), то у А.С. Пушкина *деловой* имеет ряд значений, близких к современным: ‘связанный с делом, работой, службой’, ‘знающий и опытный в делах, занятый делом, его практической стороной’, ‘относящийся к существу дела’ (ср. со словарем С.А. Кузнецова).

Примечательно и столкновение А.С. Пушкиным двух эпитетов в одном контексте, где они могут быть как синонимами: *Получили ли мои приятели письма мои* ***дельные****, т. е.* ***деловые****?* [А.С. Пушкин. Письмо П.А. Плетневу, 03.03. 1826], так и различаться нюансами значений: *Милостивый государь Николай Иванович, Благодарю Вас за Ваше письмо. Оно* ***дельное и деловое****; следовательно отвечать на него не трудно* [А.С. Пушкин. Письмо Н.И. Павлищеву, 04.05.1834]. Обсуждая пушкинское разделение писем на *деловые*, касающиеся определенных проблем и дел, и *дельные,* толковые, логичные и насыщенные структурированной информацией, можно провести параллель с актуальным делением писем на «деловые» (более официальные и выдерживающие рекомендуемый формат) и «по делу» (приятельско-коллегиальный разговор о делах).

В целом анализ эпитетов, используемых А.С. Пушкиным по отношению к письмам, позволяет лучше понять его эпистолярный стиль и особенности его восприятия переписки как средства коммуникации. В определении эпистолярных норм он опережает свое время, что отражается в семантике его эпитетов *дельный* и *деловой*, близкой к современной. При этом параллели в пушкинской и современной трактовке *дельности* и *вздора*, *дельного* и *делового,* позволяет увидеть общие черты в подходах к классификации писем в разные эпохи и указывает на то, что некоторые принципы организации эпистолярного общения остаются неизменными.

Литература

Арутюнова Н. Д. Типы значений: Оценка. Событие. Факт. М.: Наука, 1988. 341 с.

Словари русского языка // Грамота.ру [Электронный ресурс]. URL: https://gramota.ru/biblioteka/slovari (дата обращения: 04.11.2024).

У.С. Загребина
(Ижевск, Россия)
Удмуртский государственный университет
Ul.zagrebina@yamdex.ru

ИНСТРУМЕНТЫ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА И ИХ ИСПОЛЬЗОВАНИЕ В ИССЛЕДОВАНИЯХ ЛЕКСИЧЕСКОЙ СОЧЕМАЕМОСТИ ПРИЛАГАТЕЛЬНЫХ: ДИАХРОНИЧЕСКИЙ АСПЕКТ

В исследовании описаны возможности применения инструментов Национального корпуса русского языка при изучении лексической сочетаемости прилагательных в диахроническом аспекте. Рассматриваются такие функции, как «поиск коллокаций», «лексико-грамматический поиск» и «создание пользовательского подкорпуса». В ходе работы были привлечены данные панхронического и основного корпусов Национального корпуса русского языка. Сделан вывод о том, что корпус можно успешно использовать для выявления особенностей развития значений и формирования сочетаемости лексических единиц.

Ключевые слова: корпус, прилагательное, семантика, человек, диахрония, сочетаемость.

Современные автоматизированные системы во многом упрощают процесс лингвистической обработки материала, теперь в распоряжении исследователя оказываются колоссальные массивы текстов разных типов [НКРЯ]. Именно поэтому сегодня одним из активно развивающихся направлений в языкознании является корпусная лингвистика. В рамках данного направления ученые занимаются разработкой, созданием и использованием текстовых корпусов. Под корпусом, как правило, понимается «большой, представленный в электронном виде, структурированный и размеченный, филологически представительный массив языковых данных, предназначенных для решения определенных лингвистических задач» [Захаров 2005: 3].

Среди русскоязычных корпусов наиболее разработанным является Национальный корпус русского языка (далее — НКРЯ), созданный в 2003 году и включающий «тексты разных речевых сфер и разных исторических эпох — от древнерусского периода до XXI века», а значит, «на материале корпуса можно проводить диахронические

исследования» [Савчук 2019: 311]. Для нашего исследования были выбраны прилагательные с корнем -зл-.

Начать исследование следует с определения первого употребления языковых единиц, ставших предметом анализа, с этой целью мы обратились к данным панхронического корпуса, который содержит тексты древнерусского, старорусского и современного периодов развития языка, в качестве примера были выбраны прилагательные с корнем -зл-: *злой, злобный* и *злостный*. Согласно корпусным данным, самым ранним из указанных нами адъективов является *злой*, первое употребление фиксируется в текстах XI века (Изборник 1076 г., Сказание о Борисе и Глебе, Чудеса Бориса и Глеба), затем появляется *злобный* в XI–XII веках (Пчела, История Иудейской войны Иосифа Флавия, Повесть временных лет по Ипатьевскому списку), а самым поздним является слово *злостный*, которое впервые используется в текстах XVIII века (В.Н. Татищев. Произвольное и согласное разсуждение и мнение собравшегося шляхетства русского о правлении государственном). На основе полученных данных мы можем предположить, что лексическая сочетаемость прилагательного *злостный* будет уже, чем у слов *злобный* и *злой*, поскольку оно появилось в языке позже.

Чтобы получить общее представление о сочетательных возможностях прилагательных, мы воспользовались функцией «поиск коллокаций», которая доступна в основном и газетном корпусах. Для этого в качестве ключа мы ввели прилагательное, а для коллоката указали грамматический признак существительное, расстояние от – 2 до 2. Например, выбрав прилагательное *зловещий*, мы получили 252 коллокации в основном корпусе (данные на 30.10.24). Среди наиболее частотных нами выделены сочетания *зловещая тишина* (156 примеров), *зловещая тень* (103 примера), *зловещий слух* (80 примеров), *зловещий свет* (71 пример). Проанализировав все коллокации, можно сделать вывод о том, с какими существительными чаще всего сочетается прилагательное и каково его основное значение. Так, с адъективом *зловещий* чаще употребляются неодушевленные абстрактные субстантивы (*пророчество, предчувствие, призрак, молчание*), которые условно можно объединить в тематическую группу «тайна, мистика, сверхъестественное».

В НКРЯ есть функция выбора пользовательского подкорпуса, в котором исследователь имеет возможность указать сферу употребления, годы создания текстов и конкретного автора. Эту функцию можно использовать вместе с «поиском коллокаций», чтобы проследить семантическую эволюцию изучаемой языковой единицы. Например, рассмотрев нехудожественные тексты XIX века, мы нашли лишь 14 коллокаций со словом *зловещий*, а в художественных — 33. При этом показательным является то, что в текстах XX века соотношение коллокаций для указанного прилагательного значительно изменилось: мы обнаруживали 79 сочетаний в нехудожественных текстах и 116 — в художественных.

Другая удобная функция, которая есть во всех подкорпусах, называется «лексико-грамматический поиск». Она значительно облегчает поиск сочетаний различных типов. Так, создав запрос прилагательное *зловещий* с одушевленным существительным в художественной литературе XIX века, мы обнаружили лишь 2 коллокации — *зловещая птица* и *зловещий ворон*, а в текстах XX века к ним добавляются *зловещая старуха, зловещая женщина* и *зловещий человек*. Можно говорить о том, что семантика данного прилагательного становится более широкой, оно сближается по значению с прилагательным *пугающий*.

В 2023 году в НКРЯ появилась функция «портрет слова», с помощью которой можно проследить сочетаемость слова, его частотность, морфемный состав, а также корпус предлагает наиболее употребляемые однокоренные слова, похожие слова и статистику текстов. Рассмотрим на примере единицы *злой*, которая встречается чаще в художественных текстах (56%). В большинстве случаев это слово выступает как определение при неодушевленных абстрактных существительных, которые называют независящие от человека обстоятельства (воля, рок) или, наоборот, умышленное действие человека, направленное против другого человека (умысел, насмешка, шутка). Вторую группу слов чаще всего определяет прилагательное *злобный*, которое является наиболее похожим словом, согласно данным НКРЯ. Примечательно, что для прилагательного *злой* дается много похожих однокоренных слов: *озлобленный, злопамятный, презлый*. Эти единицы уточняют обобщенное значение прилагательного *злой* и выступают в более узких контекстах.

Таким образом, можно сделать вывод о том, что для проведения лингвистических исследований НКРЯ — очень ценный и удобный ресурс. Он включает тексты разных жанров и периодов развития языка, что позволяет изучить историю определенного слова. Благодаря подробной метаразметке исследователь имеет возможность выбирать контексты с изучаемой лексической единицей, задавая разные параметры (жанр, дата создания, тематика текстов и др.). Также функция «поиск коллокаций» позволяет сделать первичные выводы о сочетательных возможностях определенного слова.

Литература

Захаров В. П. Корпусная лингвистика: Учебно-метод. пособие. СПб, 2005. 48 с.

НКРЯ — Национальный корпус русского языка [Электронный ресурс]. URL: https://ruscorpora.ru/new/.

Савчук С. О. Инструментарий Национального корпуса русского языка в диахронических исследованиях // Корпусная лингвистика — 2019 Труды международной конференции. Санкт-Петербургский государственный университет; Институт лингвистических исследований РАН; Российский государственный педагогический университет им. А.И. Герцена. 2019. С. 310–316.

Н.П. Иордани
(Москва, Россия)
Институт русского языка им. В.В. Виноградова РАН
iordani.natasha@yandex.ru

ЖАНРОВЫЕ ХАРАКТЕРИСТИКИ ДЕЛОВЫХ ДОКУМЕНТОВ В СТАРОРУССКОМ КОРПУСЕ

В настоящем докладе будут рассмотрены проблемы, связанные с определением жанра деловых документов, представленных в старорусском корпусе. Сведения о жанровой принадлежности источника при его включении в состав корпуса, как правило, приводятся с опорой на издания. По этой причине при попытке задать подкорпус по параметру «жанр» в выборке могут оказаться тексты иной жанровой принадлежности, что может повлиять на результаты исследования пользователей корпуса.

Ключевые слова: Национальный корпус русского языка, старорусский корпус, старорусская деловая письменность, приказный язык, жанр делового документа.

Старорусский корпус, объем которого составляет более 9 000 000 слов, включает разнообразные тексты, созданные на великорусской территории в период с XV по XVII в. и относящиеся к разным регистрам: книжному, гибридному, деловому и бытовому. Деловой регистр в старорусском корпусе представлен многочисленными приказными документами разных жанров, которые были опубликованы в разнообразных изданиях [Сичинава 2016: 209], в том числе в дореволюционных, таких как «Русская историческая библиотека», «Акты Московского государства» и др. Каждый текст, представленный в старорусском корпусе, содержит дополнительную информацию, включающую название документа, дату создания, жанр и прочие сведения, которые заимствуются непосредственно из издания. Это дает возможность исследователю, работающему со старорусским корпусом, задать подкорпус: очертить круг текстов, отобранных по определенным параметрам, к которым относится и жанровая принадлежность текста.

В советское время проблема жанрового членения делопроизводственных источников активно изучалась. Однако в более ранних изданиях, вышедших в XIX — начале XX вв., вопросы жанрового

своеобразия текстов, относящихся к деловой письменности, не являлись центральными: главной целью такого рода публикаций было введение подобных документов в научных оборот как источников по истории России.

К таковым относятся так называемые «Донские дела», опубликованные В.Г. Дружининым в 18-ом томе серии «Русская историческая библиотека» [Дружинин 1898: 2]. Значительный объем текстов, представленных в этом издании, составляют *грамоты*, охарактеризованные издателем как *царские*. Именно эти документы сформировали в рамках старорусского корпуса подкорпус *царский грамот*, в который входит 74 текста.

Однако словосочетание *царская грамота* едва ли можно считать удачным для обозначения жанра деловых документов, поскольку оно не раскрывает специфики такого рода источников, сообщая только об адресанте документа. Так, если обратиться к текстам, представленным в этом подкорпусе, то можно отметить, что большая их часть была составлена в московских приказах от лица царя и содержала разнообразные распоряжения, адресованные должностным лицам на местах — воеводе или другим приказным людям:

(1) *Отъ царя і великого князя Михаила Федоровича всеа Русии на Донъ, въ нижние і въ верхние юрты, Донскимъ атаманомъ и казакомъ Науму Васильеву і всему Донскому Войску <…> і вы* ***бъ*** *къ намъ службу свою и радѣнье совершенно* ***показали****: по прежнему нашему указу, на сакмахъ и на рекахъ по перелазомъ і в ыныхъ въ крѣпкихъ мѣстехъ надъ ними* ***промышляли*** *всякими мѣрами неоплошно съ великимъ радѣньемъ и поискъ надъ ними* ***чинили****, сколько милосердый Богъ помочи подастъ*. [Царская грамота донским казакам с похвалою за переход нагайских мурз в холопство государя и с увещанием перезывать остальных нагайских мурз и продолжать свою службу, охраняя украйны от набегов крымцев, турок и нагайцев (1640.03.18)];

(2) *Отъ царя і великого князя Михаила Ѳедоровича всеа Русии на Донъ, въ нижние і въ верхние юрты, Донскимъ атаманомъ и казакомъ, Науму Васильеву і всему Донскому Войску <…> а къ Нагайскимъ* ***бы есте*** *мурзамъ и къ ихъ улуснымъ людемъ отъ себя* ***писали*** *и* ***посылали*** *кого пригоже по прежнему-жъ нашему*

указу, чтобъ они, помня къ себѣ наше царкое жалованье і свою прежнюю правду и шерть, были подъ нашею царскою высокою рукою по прежнему, и ***шли-бь*** *назадъ за Донъ на Нагайскую сторону*. [Царская грамота Донскому войску о сообщении крымских вестей с приказанием: идти войной на крымские улусы, если крымцы двинутся на украйные города, уговаривать нагайцев переходить в подданство к государю (1639.08.25)]

Тексты, устроенные подобным образом, в источниковедении относят к *указным грамотам*, которые были широко представлены в приказном и монастырском делопроизводстве с XV в. [Качалкин 1988: 62; Тихомиров 2003: 467].

Однако помимо *указных грамот* в этом подкорпусе встречаются источники иного содержания: они выдавались частным лицам и подтверждали их право беспрепятственного передвижения по территории государства. Подобные документы относятся к *проезжим грамотам*:

(3) *Отъ царя і великого князя Михаила Ѳедоровича всеа Русии отъ Москвы <…> до Архангельского города, и назадъ до Москвы, воеводамъ нашимъ и дьякомъ і всякимъ нашимъ приказнымъ людемъ. Били намъ челомъ Донские казаки <…> намъ бы ихъ пожаловатъ, велѣть ихъ съ Москвы отпустить къ Соловецкимъ чюдотворцомъ і велѣти бъ имъ дать нашу* ***проѣзжую грамоту***. *И по нашему указу, Донские казаки атаманъ Обакумъ Совоновъ да казаки Івашко Івановъ съ товарыщи шесть человѣкъ съ Москвы отпущены къ Соловецкимъ чюдотворцомъ*. [Царская проезжая грамота, данная атаману Аввакуму Софонову и шести казакам для отправленья на богомолье в Соловецкий монастырь (1640.01.31)]

Получается, что в подкорпусе *царских грамот* объединены два вида грамот — *указные* и *проезжие*, обладающие разным юридическим статусом и функцией. По этой причине кажется закономерным распределить эти грамоты в подкорпусы *указных* и *проезжих* грамот, которые уже существуют в рамках старорусского корпуса.

Таким образом, появление старорусского корпуса, работа с которым предполагает возможность создания выборки текстов, относящихся к одному жанру, выявило необходимость комплексного

исследования уже введенных в научный оборот текстов, опубликованных в разных изданиях в разное время, с точки зрения их жанровой принадлежности и содержания.

Литература

Дружинин В.Г. Предисловие // Русская историческая библиотека. Т. XVIII. Кн. 1. СПб., 1898. С. 1–7.

Качалкин А. Н. Жанры русского документа допетровской эпохи. Ч. 2. Филологический метод анализа документа. М.: Изд-во МГУ, 1988. 120 с.

Сичинава Д. В. Старорусские/среднерусские тексты в НКРЯ: от экстенсивной коллекции к корпусу // Textual Heritage and Information Technologies: Conference material. 2016. С. 208–210.

Тихомиров М. Н. Приказное делопроизводство в XVII в. // Тихомиров М. Н. Российское государство в XV–XVII вв. М.: Языки славянской культуры, 2003. С. 445–496.

И.Б. Качинская
(Москва, Россия)
МГУ имени М.В. Ломоносова
kacza@yandex.ru

ДИАЛЕКТНЫЙ КОРПУС: ПРОБЛЕМЫ ОТБОРА ИСТОЧНИКОВ, ЛИНГВИСТИЧЕСКОЙ И ЭКСТРАЛИНГВИСТИЧЕСКОЙ РАЗМЕТКИ[1]

Диалектный корпус пополняется главным образом за счет уже изданных хрестоматий, а также за счет передаваемых в Корпус неопубликованных полевых экспедиционных материалов. За последние 5 лет в подкорпус переданы коллекции более чем из 15 регионов, многие тексты имеют аудиосопровождение, есть несколько видео. Возникли проблемы пополнения портфеля из-за резкого снижения количества диалектологических экспедиций и появления новых диалектных корпусов. Требуется совершенствование разметки, главным образом экстралингвистической.

Ключевые слова: русская диалектология, корпусная лингвистика, национальный корпус русского языка

1. За последние пять лет Диалектный корпус НКРЯ пополнился материалами из Архангельской, Волгоградской, Вологодской, Кировской, Нижегородской, Пермской, Псковской, Смоленской, Тверской, Томской, Тюменской областей, включены материалы русских говоров Башкирии, Забайкалья, Нижнего Поволжья, Белоруссии и Азербайджана.

243 текста имеют аудиосопровождение. Это тексты из Волгоградской, Тверской, Псковской, Смоленской областей, большая коллекция представлена из островного русского говора Азербайджана — из деревни Ивановки, где уже более 200 лет проживают русские протестанты, первоначально молокане, а в настоящее время уже не только молокане, но и баптисты, и пятидесятники.

[1] Работа выполнялась благодаря поддержке гранта Министерства науки и высшего образования № 075-15-2020-793 «Компьютерно-лингвистическая платформа нового поколения для цифровой документации русского языка: инфраструктура, ресурсы, научные исследования».

4 текста имеют видеосопровождение: из Архангельской обл. и из Ивановки (Азербайджан). В ближайшие годы будет увеличено количество видеоматериалов.

Диалектный корпус пополняется главным образом за счет уже изданных региональных хрестоматий, которые, как правило, выходили ничтожным тиражом в качестве пособий по диалектологии для студентов местных вузов, а также за счет передаваемых в Корпус неопубликованных полевых экспедиционных материалов.

В последнее время приток полевых материалов в Диалектный подкорпус НКРЯ резко сократился. Причин несколько.

Во-первых, со времени пандемии во многих вузах прекратили финансировать практику по диалектологии, вывели ее за пределы учебных программ. Повсеместно сокращается количество часов по русской диалектологии, а где-то и вовсе эту дисциплину отменили.

Во-вторых, многие вузы, в которых диалектологические экспедиции проводятся (или проводились) регулярно, создают свои собственные корпуса: Саратовский, Екатеринбурский, Томский, Петрозаводский, Тамбовский и нек. др. Часто это закрытые корпуса, т.е. не имеющие выхода в интернет. Если выход в интернет появляется, то может потом на несколько лет исчезнуть, как это происходило со «Звучащей хрестоматией тамбовских говоров», которая то появлялась на просторах интернета, то исчезала: у вуза, по-видимому, не было возможности — как технической, так и материальной, — держать ее на своем сайте. В настоящее время Тамбовская хрестоматия, поменяв несколько адресов, расположилась на сайте Тамбовской областной универсальной научной библиотеки им. А.С. Пушкина [Тамбовская хрестоматия].

Часто местные корпуса имеют лексикографическую или идеографическую, тематическую направленность, то есть там нет задачи выставлять собственно тексты.

Большая коллекция диалектных текстов в сопровождении аудионарезки находится в свободном доступе на сайте Лаборатории языковой конвергенции НИУ ВШЭ [ВШЭ].

Иногда в хрестоматиях присутствуют лингвистические, исторические, культурные комментарии, и хорошо было бы продумать возможность их размещения на платформе Диалектного корпуса.

2. Отбор источников. В Диалектный корпус НКРЯ попадают преимущественно нарративы, т.е. связные тексты, которые могут представлять интерес не только для лингвистов, но и для историков, культурологов, психологов, фольклористов. Не попадает нарезка, характерная для региональных словарей или тематических исследований, пересказ (позднее воссоздание) текстов или искусственно созданные беллетризованные тексты на «диалектном языке», не включается также собственно беллетристика, даже если она наполнена диалектизмами.

Аутентичные тексты принимаются как в транскрипции, так и в орфографизированном и даже в орфографическом исполнении, желательно с сохранением ударений и особенностей грамматики. Поощряется подача текстов с аудиосопровождением.

В последние два десятилетия (фактически с начала XXI века) все экспедиции привозят материалы в виде аудио и видео. Однако до наступления цифровой эры с записывающей техникой были большие сложности.

Сегодня основная сложность — это расшифровка текстов. Многие держатели были бы готовы поделиться своими коллекциями аудио. Но некому расшифровывать материалы. Мы отчаянно нуждаемся в появлении хороших программ распознавания диалектных текстов, эти программы уже появляются, качество их растет, но после них требуется существенная доработка.

3. Лингвистическая разметка. Особенностью Диалектного корпуса НКРЯ является не только характерная для всего Корпуса грамматическая разметка, но и возможность искать слово сразу по многим подкорпусам, по параметрам, которые одинаково заданы для всего Корпуса. В метаразметке указываются некоторые фонетические особенности: противопоставленные (оканье ~ аканье, заднеязычный звонкий ~ фрикативный) и непротивопоставленные (рефлексы ятя, /а/, цоканье). Поверх стандартной грамматической разметки осуществляется специфическая диалектная. Во многих случаях указывается семантика, особенно если в хрестоматии имеется словарик диалектных слов.

4. Экстралингвистическая разметка содержит адрес-сопровождение к каждому тексту (название текста, где он был записан,

кем, от кого) и метатекстовую разметку, к которой относится указание на жанр текста, его тематику, место и время описываемых событий. Раздел по тематике необходимо изменить. Когда-то он создавался на основе вопросников для Лексического атласа народных говоров [ЛАРНГ]. Для нас это оказалось слишком дробное членение. Подготовлен более приемлемый тематический список, но, к сожалению, пока нет технической возможности внести эти изменения в специально созданную для подготовки Диалектного корпуса программу «Рабочее место диалектолога».

Литература

ВШЭ — Лаборатории языковой конвергенции НИУ ВШЭ [Электронный ресурс]. URL: https://lingconlab.ru/

ЛАРНГ — Программа собирания сведений для Лексического атласа русских народных говоров. [Электронный ресурс]. URL: https://iling.spb.ru/departments/larng

Тамбовская хрестоматия — Звучащая хрестоматия тамбовских говоров [Электронный ресурс]. URL: https://elibrary.tambovlib.ru/?ebook=9711

Е. В. Кашкин, И. А. Хомченкова
(Москва, Россия)
Институт русского языка им. В. В. Виноградова РАН
egorka1988@gmail.com, irina.khomchenkova@yandex.ru

КРУГЛЫЙ СТОЛ «КОРПУСНЫЕ МЕТОДЫ В ИССЛЕДОВАНИИ ЯЗЫКОВЫХ КОНТАКТОВ»

В работе представлена ключевая информация о круглом столе «Корпусные методы в исследовании языковых контактов». Он посвящен корпусам нестандартных вариантов русского языка, контактирующих с иными языками; корпусам некоторых миноритарных языков России и возможностям их использования для анализа контактно обусловленных явлений (в том числе в изучении переключения кодов); параллельным корпусам.

Ключевые слова: корпусная лингвистика, языковые контакты, миноритарные языки, полевая лингвистика, социолингвистика, переключение кодов.

1. Постановка проблемы

Круглый стол, организованный авторами данной работы, посвящен использованию корпусов при исследовании языковых контактов: методологическим и техническим аспектам подготовки корпусов, служащих инструментом таких исследований, подходам к разметке, примерам анализа конкретных явлений.

Тематика языковых контактов приобретает всё бóльшую важность, находясь на стыке теоретической лингвистики и многих других областей (социологии, культурологии, истории, географии и др.) и вписываясь тем самым в современный междисциплинарный контекст, см., например, [Bullock, Toribio (eds.) 2009; Bhatia, Ritchie (eds.) 2013; Grant (ed.) 2019; Matras 2020]. Исследовательские вопросы состоят в том, какие явления и модели чаще, а какие реже подвергаются заимствованию, чем объясняются выявленные закономерности, какие теоретические и когнитивные механизмы скрываются за контактно обусловленными процессами. Изучение вариантов русского языка, развившихся во взаимодействии с другими языками, дополняет эмпирическую картину того, как в принципе может быть устроен русский язык и каковы пределы его внутреннего варьирования.

Важность корпусного метода обусловлена в данном случае тем, что исследования языковых контактов зачастую ведутся методами наблюдения или анкетирования без четко структурированных корпусных данных (ср. традиционные исследования «интерференции» и ее «преодоления» в русском языке), или же корпуса выполняют вспомогательную роль (например, исследователь может найти в обычном корпусе ту или иную конструкцию, а далее с не всегда достаточно эксплицированной аргументацией трактовать ее как появившуюся под контактным влиянием).

Некоторые контактные явления (например, заимствования, переключение кодов) можно исследовать по корпусам, создаваемым в рамках документационных проектов. При этом требует решения вопрос того, как отражать в корпусах такие явления оптимальным для пользователя образом. Кроме того, в последние десятилетия были разработаны корпуса, нацеленные на систематизацию контактно обусловленных явлений (см. раздел 2). При создании этих корпусов принимались различные решения о подборе материала и о принципах его разметки; ср., например, наборы тегов (при их наличии) для нестандартных явлений. Обмен опытом в этой сфере полезен как для совершенствования разработки подобных ресурсов, так и для поиска общих закономерностей и различных черт в лингвистических характеристиках нестандартных вариантов русского языка.

2. Доклады

Программа круглого стола включает пять докладов. Выступление **Е. В. Кашкина** посвящено корпусным ресурсам ИРЯ РАН, предназначенным для изучения русского языка в контактной и типологической перспективах. Во-первых, это корпус русской речи носителей языков России [Ruscontact], см. также [Khomchenkova et al. 2019]. Он включает образцы устной русской речи носителей энецкого, нганасанского, ненецкого, нанайского, ульчского, эвенского, чукотского языков. В настоящее время в корпус интегрируется коллекция русских текстов, записанных от носителей горномарийского языка; эта работа отдельно обсуждается в докладе. В корпусе представлена детализированная разметка языковых явлений, предположительно имеющих контактную природу: нестандартное употребление фонетических единиц (например, озвончение/ оглушение), морфологических форм (числа, вида и др.), синтакси-

ческих конструкций и дискурсивных единиц (сбои согласования и управления, особое использование дискурсивного маркера и др.), лексические кальки.

Во-вторых, обсуждаются параллельные корпуса, размещенные на основной платформе НКРЯ (см. о них [Сичинава 2019]). В них размещены тексты разнообразных жанров на литературном русском языке и переводные (в одном или другом направлении) соответствия на других языках. На данный момент в параллельных корпусах обработаны данные более чем 30 языков — как крупнейших языков мира (английского, испанского, хинди, японского и др.), так и языков, носители которых активно контактируют с русским языком в той или иной социолингвистической ситуации (армянского, башкирского, вепсского, хакасского и др.). Использование этого материала дает возможность сопоставительно-типологических исследований русского языка, а также анализа контактных процессов.

Доклад **И. А. Хомченковой** посвящен контактным явлениям, которые можно исследовать с помощью корпусов миноритарных языков, — на материале корпусов горномарийского языка [КГМ] и татышлинского говора удмуртского языка [КТУ]. В качестве иллюстраций рассматривается несколько результатов контактов между соответствующими идиомами (реципиентами) и русским языком (донором). Во-первых, корпуса позволяют узнать, какие русские элементы используются в горномарийской и удмуртской речи (и сравнить результаты, например, с иерархиями заимствуемости по [Matras 2007]). Во-вторых, можно не только оценить их абсолютную частотность, но и сравнить с аналогами (при их наличии) в языке-реципиенте. В докладе внимание фокусируется на русских числительных, союзах и частицах. Показано, что не все иерархии по [Matras 2007] подтверждаются на проанализированном материале. Также обсуждаются ограничения на использование русских элементов с точки зрения (ин)конгруэнтности конструкций языков-реципиентов и языка-донора. Например, условные конструкции и в русском, и в горномарийском включают финитный глагол, однако позиция союзов различается (препозитивный в русском, постпозитивный в горномарийском). Такая неполная конгруэнтность позволяет совмещать структуры двух языков, в результате чего может происходить дублирование союза (*если... гӹнь*).

Тематика доклада **Г. А. Мороза** — исследование вариативности в устных корпусах Международной лаборатории языковой конвергенции НИУ ВШЭ (см. сайт [Lingconlab]). Коллектив лаборатории в сотрудничестве с лингвистами из других организаций ведет разработку корпусов ряда малых языков России (абазинского, башкирского, даргинских языков и др.), русских диалектов, корпусов билингвального русского (варианты русского языка, находящиеся в контакте с башкирским, луговым марийским, чувашским, хантыйским и нек. др. языками), других электронных ресурсов. На этом материале с применением статистических методов исследуются, в частности, особенности русской речи с учетом контактно обусловленных, внутриязыковых, социолингвистических факторов, например, нестандартное опущение предлогов [Panova, Philippova 2021; Яковлева (в печати)], отличный от литературного русского синтаксис количественных конструкций [Naccarato, Moroz 2023], ряд других морфосинтаксических явлений.

Выступление **Е. В. Рахилиной** и **А. К. Казкеновой** посвящено еще одному ресурсу, развиваемому в НИУ ВШЭ, — Русскому учебному корпусу [RLC], см., например, [Rakhilina et al. 2016]. Корпус содержит русские тексты, созданные людьми, которые не владеют русским языком как родным и изучают его либо являются эритажными говорящими (т. е. начали усваивать русский в детстве как первый язык и затем, как правило, в результате эмиграции перешли на другой доминантный язык). Как и в упомянутом выше корпусе [Ruscontact], в этом корпусе также размечаются явления, нехарактерные для стандартного русского языка. В качестве иллюстраций и аргументации ряда теоретических выводов авторы представляют результаты работы с казахским подкорпусом Русского учебного корпуса, см., например, [Рахилина, Казкенова 2020; Рахилина и др. 2021; Казкенова, Рахилина 2024].

Н. М. Стойнова рассматривает в своем докладе подходы к отражению переключения кодов в устных корпусах малых языков (ср., например, сочетание лексических и грамматических фрагментов русского и горномарийского языков во взятом из естественного текста предложении *Тӹ сельскохозяйственный техникумышкыжы мӹньӹ после седьмого класса даже поступаенäм ылы, по конкурсу не прошла* ‘В тот сельскохозяйственный техникум я после седьмого класса даже

поступала, по конкурсу не прошла’). Обсуждаются проблемы документации, транскрибирования, лингвистической разметки подобного материала с опорой, в частности, на опыт разработки корпуса с переключением кодов в ИРЯ РАН под руководством автора доклада [CS corpus] и на опыт корпусов Гамбургского проекта [INEL].

Литература

Казкенова А. К., Рахилина Е. В. От глагольных форм — к служебным словам: грамматикализация в русском и казахском языках сквозь призму текстов билингвов // Russian Linguistics. 2024. № 48 (16). С. 1–23.

КГМ — Корпус горномарийского языка [Электронный ресурс]. URL: https://hillmari-exp.tilda.ws/corpus (дата обращения: 03.12.2024).

КТУ — Корпус татышлинского удмуртского [Электронный ресурс]. URL: https://udmurt.web-corpora.net/tatyshly/ (дата обращения: 03.12.2024).

Рахилина Е. В., Казкенова А. К. Маркирование итератива в русской речи носителей казахского языка // Труды Института русского языка им. В. В. Виноградова. 2020. №4. С. 168 — 192.

Рахилина Е. В., Казкенова А. К., Ахапкина Я. Э. После, через, спустя во временны́х контекстах: из наблюдений над текстами казахско-русских билингвов // Вестник Томского государственного университета. Филология. 2021. № 73. С. 93–113.

Сичинава Д. В. Параллельные тексты в составе Национального корпуса русского языка: новые языки и новые задачи // Труды Института русского языка им. В. В. Виноградова. 2019. № 3. С. 41–61.

Яковлева А. В. Опущение предлогов в русской речи марийских и бесермянских билингвов: исследование на основе устных корпусов // Труды института русского языка им. В. В. Виноградова. В печати.

Bhatia T., Ritchie W. (eds.). The handbook of bilingualism and multilingualism. Oxford: Wiley Blackwell, 2013. 976 pp.

Bullock B., Toribio A. (eds.). 2009. The Cambridge handbook of linguistic code-switching. Cambridge: Cambridge University Press. 422 pp.

CS corpus — Корпус устных текстов с разметкой переключения кодов [Электронный ресурс]. URL: http://web-corpora.net/ruscontact/CS.html (дата обращения: 03.12.2024).

Grant A. (ed.). The Oxford handbook of language contact. Oxford: Oxford University Press, 2019. xxix + 757 pp.

INEL — INEL Ressourcen Portal [Online resource]. URL: https://inel.corpora.uni-hamburg.de/portal/#en (date of access: 03.12.2024).

Khomchenkova I., Pleshak P., Stoynova N. The corpus of contact-influenced Russian of Northern Siberia and the Russian far East // Computational linguistics and intellectual technologies: Proceedings of the international conference “Dialogue 2019”. M.: RSUH, 2019. P. 253–264.

Lingconlab — Ресурсы Международной лаборатории языковой конвергенции НИУ ВШЭ [Электронный ресурс]. URL: https:// lingconlab.ru/ (дата обращения: 03.12.2024).

Matras Y. The borrowability of grammatical categories // Grammatical borrowing in cross-linguistic perspective / Y. Matras and J. Sakel (eds.). Berlin, NY: Mouton de Gruyter, 2007. P. 31–74.

Matras Y. Language contact. Cambridge: Cambridge University Press, 2020. xix + 409 pp.

Naccarato Ch., Moroz G. Non-standard numeral constructions in L2 Russian: A corpus-based study // Talk at the conference "Indigenous languages of Russia in contact with Russian" (Moscow, Vinogradov Russian Language Institute of the RAS, 15–16 September 2023).

Panova A., Philippova T. When a cross-linguistic tendency marries incomplete acquisition: Preposition drop in Russian spoken in Daghestan // International journal of bilingualism. 2021. № 25(3). P. 640–667.

Rakhilina E., Vyrenkova A., Mustakimova E., Ladygina A., Smirnov I. Building a learner corpus for Russian // Proceedings of the joint workshop on NLP for computer assisted language learning and NLP for language acquisition at SLTC, Umeå, 16th November 2016. [Online resource]. URL: http://aclweb.org/anthology/W16-65 (date of access: 03.12.2024).

RLC — Русский учебный корпус [Электронный ресурс]. URL: http://web-corpora.net/RLC/ (дата обращения: 03.12.2024).

Ruscontact — Корпус русской речи носителей языков России [Электронный ресурс]. URL: http://web-corpora.net/ruscontact/ corpus.html (дата обращения: 03.12.2024).

Н. И. Киреев
(Париж, Франция)
Высшая нормальная школа Парижа
nkireyev@yandex.ru

НКРЯ И ИСТОРИЧЕСКАЯ АКЦЕНТОЛОГИЯ

В докладе пойдёт речь о том, как можно использовать НКРЯ для исследований по истории русского ударения; в первую очередь — о силлаботонической поэзии в составе поэтического подкорпуса. После обзора существующих исследований и подходов в этой области будут продемонстрированы основные сложности в использовании поэтических данных в исторической акцентологии и возможные пути их решения.

Ключевые слова: историческая акцентология, поэтический подкорпус, просодия, русская акцентология, силлаботоническое стихосложение, стиховедение.

После введения Петром I в 1708 году гражданского шрифта надстрочные знаки, в частности, отмечающие место ударения, стали печататься лишь в церковных текстах. Старопечатные же книги содержали более-менее систематическую акцентуацию, поэтому они являются важным источником для исторической акцентологии русского языка (см., например [Зализняк 2019]).

После 1708 года знаки ударения, помимо церковнославянских текстов, сохраняются также в словарях, но далеко не во всех, и не все лексикографические данные достаточно надёжны. В этом контексте особо важным оказывается появление в 1740-х годах силлаботонической поэзии на русском языке.

Силлаботоника основана на упорядочивании ударных и безударных слогов в стихе. Несмотря на расхожее обывательское представление о том, что русская поэзия, особенно старая, содержит массу licentiae poeticae, исследования показывают, что почти все неожиданные для современного читателя ударения, которые встречаются в текстах русских классиков, отражают лингвистическую реальность своего времени. Поэтому для любых исследований, касающихся истории русского ударения XVIII—XX веков, данные русской поэзии представляют собою большую ценность.

Легко заметить, что образцовые исследования в этой области, выполненные в докорпусную эпоху (такие, как [Воронцова 1979; Еськова 2008], отчасти также [Булаховский 1948]), выполнены на

материале классической поэзии фактически по корпусной методологии (или, если угодно, в рамках корпусной идеологии [Плунгян 2008]).

Создатели НКРЯ изначально продумывали возможности использования корпуса для исследований по исторической акцентологии (см., например, [Гришина 2009]). С тех пор был опубликован ряд методологических обзоров (отметим из последних [Корчагин 2019, Орехов, Савчук 2019]), появилось нескольких блестящих исследований (укажем [Сичинава 2014, 2015]). Ясно, что возможности в этой области простираются даже дальше, чем изучение истории ударения конкретного слова или группы слов, а позволяют строить и проверять более сложные акцентологические гипотезы [Piperski, Kukhto 2016; Piperski, Kukhto 2021].

Тем не менее, исследований русского ударения на основе русской силлаботонической поэзии не так много, и по проницательному наблюдению Д. В. Сичинавы, славянская историческая акцентология и славянское стиховедение являются «малознакомыми ровесниками» [Сичинава 2020].

В докладе мы постараемся осветить возможные применения НКРЯ для изучения истории русского ударения. Заметим, что поэтический подкорпус позволяет не только наблюдать смену места ударения в том или ином слове, но и, например, исследовать процесс клитизации, т. е. утраты словом собственного ударения (попытку такого исследования на примере слова *или* см. в [Киреев 2022]). Будет сказано о главных проблемах в интерпретации данных классической поэзии и намечены пути их разрешения.

Кроме того, кратко будет рассмотрен и диахронический аспект других модулей НКРЯ, имеющих акцентологическую разметку (устного, мультимедиа, наивной поэзии и др.).

Литература

Булаховский Л. А. Русский литературный язык первой половины XIX века. Фонетика. Морфология. Ударение. Синтаксис. Киев: Радянська школа, 1948.

Воронцова В. Л. Русское литературное ударение XVIII–XX веков. М.: Наука, 1979. 328 с.

Гришина Е. А. Корпус «История русского ударения» // В. А. Плунгян, Е. В. Рахилина, Т. И. Резникова (ред.). Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб: Нестор-История, 2009. С. 150–174.

Еськова Н. А. Нормы русского литературного языка XVIII–XIX веков: Ударение. Грамматические формы. Варианты слов. Словарь. Пояснительные статьи. М.: Рукописные памятники Древней Руси, 2008. 960 с. (Studia philologica).

Зализняк А. А. Древнерусское ударение: Общие сведения и словарь. 2-е изд., расширенное и переработанное. М.: Издательский дом ЯСК, 2019.

Киреев Н. И. Акцентологическая история слова *или* в XVII–XX веках: корпусные данные //

Труды Института русского языка им. В. В. Виноградова РАН. 2022. № 3 (33). С. 162—180.

Корчагин К. М. Зачем нужен поэтический корпус и как его использовать // Русская речь. 2019. № 6. С. 113–127.

Орехов Б. В., Савчук С. О. Акцентологический корпус как инструмент для исследования русского ударения // Труды Института русского языка им. В. В. Виноградова. Вып. 21. М., 2019. С. 61–82.

Плунгян В. А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. 2008. № 16 (2). С. 7–20.

Сичинава Д. В. Акцентуация глагола *быть* в русском стихе // В. А. Плунгян, Л. Л. Шестакова (ред.). Корпусный анализ русского стиха. Вып. 2. М.: Азбуковник, 2014. С. 48–74.

Сичинава Д. В. Братец и *сударь*: энклитические обращения в стихе Грибоедова // Язык, литература, культура: Актуальные проблемы изучения и преподавания. Вып. 11. М.: Макс Пресс, 2015. С. 148–157.

Сичинава Д. В. Русская историческая акцентология и стиховедение: малознакомые ровесники // Кибрик А. А. и др. (ред.). ВАПросы языкознания: Мегасборник наностатей. Сб. ст. к юбилею В. А. Плунгяна. М.: Буки-Веди, 2020. С. 111–115.

Piperski A. Ch., Kukhto A. V. Intra-speaker stress variation in Russian: A corpus-driven study of Russian poetry // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 1–4 июня 2016 г.). Вып. 15 (22). М.: Изд-во РГГУ, 2016. С. 540–550.

Piperski A., Kukhto A. Inferring stress placement variability from a poetic corpus // Journal of Slavic Linguistics. 2021. Vol. 29, extra issue: Formal Approaches to Slavic Linguistics 28 Proceedings. P. 1–15.

О. А. Козан
(Анкара, Турция)
Университет имени Хаджи Байрама Вели
olena.kozan@hbv.edu.tr

НКРЯ В ПРОЦЕССЕ ПРОФЕССИОНАЛЬНОЙ ПОДГОТОВКИ ФИЛОЛОГА И ПЕРЕВОДЧИКА (НА ПРИМЕРЕ ТУРЕЦКО-РУССКОЙ ЯЗЫКОВОЙ ПАРЫ)

В данной работе представлен опыт использования Национального корпуса русского языка (НКРЯ) в процессе профессиональной подготовки филологов и переводчиков турецко-русской языковой пары на кафедре русского языка Анкарского университета имени Хаджи Байрама Вели. В докладе подчеркивается и обосновывается идея о том, что для профессиональной работы с текстом на русском языке, созданным человеком или генеративной моделью, у будущего специалиста должна быть выработана семантическая осознанность по отношению к языковым единицам и их связям, которую можно развивать, используя возможности НКРЯ.

Ключевые слова: корпус, НКРЯ, перевод, турецкий, коллокации

Современные реалии, формируемые процессом стремительного развития технологий, прежде всего генеративных моделей, создающих тексты любой природы, заставляют задуматься над траекторией дальнейшего развития гуманитарных наук, в частности, филологии и переводоведения. На протяжении истории текст являлся и объектом изучения, и «продуктом» этих областей гуманитарного знания, причем единственным создателем текстов — в самом широком понимании — выступал человек. С появлением генеративных моделей, обученных на больших языковых данных, Homo Sapiens потерял власть над текстом, более того — власть над «высшей формой речи», по мнению Иосифа Бродского — поэзией: «Если тем, что отличает нас от прочих представителей животного царства, является речь, то литература, и в частности, поэзия, будучи высшей формой словесности, представляет собою, грубо говоря, нашу видовую цель» [Бродский 1987]. И если это отличие до сих пор действительно в отношении «животного царства», то с точки зрения дихотомии естественный — искусственный интеллект оно нивелируется. В данном контексте возникает много вопросов, а именно: какими знаниями и навыками должен обладать филолог и/или переводчик — а в более широком смысле и

гуманитарий — будущего? Как вообще должны развиваться филология и переводоведение в рамках «цифровых гуманитарных наук» [Антопольский и др. 2023]? Эти вопросы послужили отправной точкой для формирования «дорожной карты» по развитию семантической осознанности у будущего специалиста по русскому языку в условиях отсутствия языковой среды на примере турецко-русской языковой пары в Турции. Под семантической осознанностью в рамках данной работы подразумеваются знания и навыки, позволяющие специалисту (носителю турецкого языка) работать с созданием текста на иностранном языке (языке перевода), в нашем случае речь идет о русском языке. Принимая во внимание новые реалии, стоит отметить, что специалист по языку — как бы он ни назывался в будущем — столкнется (и уже сталкивается) не только с процессом создания текста на иностранном языке, но и с готовыми «текстовыми продуктами» (текстоидами? [Боронин 2016]), причем произведенными как человеком, так и машиной. В данном контексте основную задачу в процессе профессиональной подготовки специалиста по языку можно сформулировать как формирование навыков анализа готового текста, в том числе и перевода (постпереводческий анализ), на иностранном языке. Реализация данной задачи представляется возможной только при обращении к НКРЯ и его возможностям [Савчук и др. 2024].

На кафедре русского языка Анкарского университета имени Хаджи Байрама Вели ведутся работы по созданию учебных материалов для турецко-русской языковой пары в рамках поставленных задач. Прежде всего речь идет о выработке у студентов навыка анализа готового переводческого решения на основе информации о языковых единицах, которую они могут получить, используя возможности НКРЯ. В качестве примера разработанного учебного материала можно привести пособие «Перевод в эпоху искусственного интеллекта. Анализ переводов человека и машины на основе корпуса» [Kaşoğlu, Kozan 2023]. Данное пособие стало «продуктом» работы магистрантов кафедры русского языка в рамках учебного курса «Постпереводческий анализ». Материалом для пособия послужили тексты на актуальные темы, касающиеся общества, экономики, политики и других областей, отобранные с ведущих информационных порталов на турецком языке (исходный текст). На основании предпереводческого анализа исходного текста студентами был предложен перевод без исполь-

зования возможностей генеративных моделей (перевод-1 на русский носителем турецкого языка). Перевод-2 являлся машинным переводом. А перевод-3, представленный в материалах, являлся готовым переводом, опубликованным на платформе www.inosmi.ru (перевод, принимаемый в качестве эквивалентного и адекватного варианта, выполненного и/или отредактированного носителем русского языка). Задача студентов состояла в постпереводческом анализе и обосновании своих выводов об ошибках и их потенциальных причинах, а также в выработке «алгоритма» анализа переводческого решения на основе данных, которые можно получить, обратившись к НКРЯ. В рамках данного доклада представлен алгоритм работы с корпусом на одном из примеров, описанных в пособии.

Литература

Антопольский А. Б., Бонч-Осмоловская А. А., Бородкин Л. И. Цифровые гуманитарные исследования. Красноярск: СФУ, 2023. 272 с.

Боронин А. А. К вопросу о текстоидах // Вопросы современной лингвистики. 2016. №2. С. 26–31.

Бродский И. Нобелевская лекция [Электронный ресурс]. URL: https://lib.ru/BRODSKIJ/lect.txt_with-big-pictures.html

Савчук С. О., Архангельский Т. А., Бонч-Осмоловская А. А., Донина О. В., Кузнецова Ю. Н., Ляшевская О. Н., Орехов Б. В., Подрядчикова М. В. Национальный корпус русского языка 2.0: новые возможности и перспективы развития // Вопросы языкознания. 2024. №2. С. 7–34.

Kaşoğlu A., Kozan O. Yapay Zekâ Çağında Çevirmen Olmak. Derlem Tabanlı İnsan ve Makine Çevirisi İncelemeleri (Türkçe-Rusça Örneğinde). Çanakkale: Paradigma. 2023. 113 p.

Д. В. Колесова, О. О. Лисова, Т. И. Попова, С. В. Чигинцева
(Санкт-Петербург, Россия)
Санкт-Петербургский государственный университет
d.kolesova@spbu.ru, o.lisova@spbu.ru, t.popova@spbu.ru,
st076678@student.spbu.ru

ЯЗЫКОВЫЕ МАРКЕРЫ ДЛЯ ПОИСКА ОПРЕДЕЛЕННОГО ТИПА ТЕКСТА В НКРЯ (НА ПРИМЕРЕ ПОИСКА ТЕКСТОВ ПРАКТИЧЕСКОГО РАССУЖДЕНИЯ)

НКРЯ открывает большие возможности для проведения исследований, а также для повышения эффективности работы преподавателя русского языка. Однако в настоящее время у исследователя и преподавателя нет возможности найти в корпусе текстовые фрагменты, которые можно отнести к тем или иным типам текста. Результат поиска по НКРЯ примеров практического рассуждения может быть полезен для выявления языковых единиц и прагматических характеристик, приводящих к составлению картотеки примеров типов текста.

Ключевые слова: тип текста, практическое рассуждение, корпусное исследование

Национальный корпус русского языка (НКРЯ) включает разнообразные типы текстов, что обеспечивает его репрезентативность и функциональность для различных лингвистических исследований. Основные категории текстов в НКРЯ следующие: 1. Прозаические тексты (художественные, мемуарно-биографические, публицистические, научные и учебные). 2. Современные художественные произведения: охватывают различные жанры и направления. 3. Устная речь (публичная и частная). 4. Поэтические тексты. 5. Диалектные тексты. 6. Параллельные тексты (тексты на русском языке в сопоставлении с их переводами на другие языки). 7. Специальные корпусы (корпус акцентологии и газетный корпус). Эти подкорпусы позволяют исследовать язык в его многообразии, включая как письменные, так и устные формы, что делает НКРЯ ценным инструментом для лингвистов и филологов.

Однако тем, кто преподает русский язык, довольно сложно воспользоваться существующим НКРЯ в своей профессиональной и исследовательской деятельности, поскольку в корпусе сейчас нет возможности задать собственно текстовые критерии.

В рамках изучения такого текста, как практическое рассуждение, мы попробовали обратиться за примерами к поиску в НКРЯ.

Кратко представим, что лингвистике известно про этот тип текста. Практическое рассуждение (ПР) — это языковое воплощение практического мышления, которое необходимо для того, чтобы совершилось действие [Арутюнова 1987: 5], целью ПР является анализ соответствующей проблемной ситуации, поиск выхода из нее, а также прогнозирование развития событий [Коньков 1995: 55–58]. Разумеется, ПР чаще всего связано с повседневной жизнью человека, именно в этой сфере жизни постоянно происходят действия, а субъект всегда непосредственно включен в представленную практическим рассуж дением ситуацию. Для отнесения текста к ПР мы опираемся в первую очередь на его информационную структуру (суждение, решение, действие), но полагаем, что также должны учитываться следующие параметры: устный/письменный, форма речи (монолог/диалог), сфера общения (бытовая, бытийная, социокультурная, рекламная и т.д.), тематика (ПР в бытовой и бытийной сферах будет существенно различаться как в проблемной ситуации, так и в части аргументов, лежащих в основе решения, и самом действии).

Анализ отобранных текстов, которые были оценены авторами проекта как практическое рассуждение, привел нас к предположению, что возможно выделить некоторые связки лексических единиц, которые могут быть заданы в поиске НКРЯ для нахождения примеров текстов.

Мы вводили следующие маркеры ПР: «превосходная мысль», «что если», «что если я буду», «решить», «лучше я», «отсюда вытекает», «спрашивается», «вопрос в том», «если… то», «во-первых … во-вторых», «чем… тем», «я долго думал(а)», «следовательно» и др.

Было отмечено, что лучшие результаты даёт маркер «если… то» в комбинации с другими маркерами, в особенности — с союзом «чем… тем», а также комбинация «во-первых… во-вторых». Среди выводимых текстов есть примеры рассуждения, в том числе — ПР, поиск которого и является нашей задачей. Отметим, что при лексико-грамматическом поиске в НКРЯ для корректного выведения примеров текстов ПР расстояние между всеми словами в указанных маркерах должно быть не менее 10 слов. Дальнейшей задачей преподавателя-исследователя является выбор текста, который соответствует уровню группы и теме урока.

Полагаем, что подкорпус, позволяющий среди великого множества имеющихся в корпусе текстов отобрать материал в рамках

функционально-смысловых типов речи, был бы очень востребован среди русистов, как теоретиков, так и преподавателей. Специфика языкового воплощения описания, повествования и рассуждения уже изучена и описана [Функционально-смысловые единицы речи, 2017], так что представляется возможным выделить собственно языковые маркеры принадлежности текста к тому или иному типу речи. Проверка степени эффективности использования того или иного подобного языкового маркера — задача, которая требует времени и экспертной оценки, однако представляется вполне возможной.

Кроме того, важно учитывать жанровую принадлежность текстов [Дементьев, 2020]. Частично жанры учитываются в существу ющих подкорпусах, однако указание жанра приводит не к поиску текстов, а к составлению подкорпуса из текстов того или иного жанра).

Полагаем, что выявление критериев для поиска типов текста в НКРЯ, проверка их применимости к исследовательской и преподавательской деятельности — актуальная лингвистическая задача. Если такие критерии будут выявлены и представлены таким образом, чтобы ими мог пользоваться каждый пользователь НКРЯ, то это обеспечит более глубокое понимание функционально-смысловых типов речи и специфики их функционирования в русском дискурсе.

Литература

Арутюнова Н. Д. Практическое рассуждение и язык // Сущность, развитие и функции языка: [Сб. ст.] / Отв. ред. Г. В. Степанов. М.: Наука, 1987.

Дементьев В. В. Что дало жанроведение современной лингвистике? // Жанры речи. 2020. № 3 (27). С. 172–194. DOI: https://doi.org/10.18500/23110740202032717219 4

Коньков В. И. Речевая структура газетного текста. СПб. : Изд-во С.-Петербургского ун-та, 1995. 158 с.

Функционально-смысловые единицы речи: типология, исходные модели и принципы развертывания /Под общ. ред. К. А. Роговой. СПб: Златоуст, 2017. 320 с.

М. В. Копотев
(Хельсинки, Финляндия)
Хельсинкский университет
mihail.kopotev@helsinki.fi

КОРПУСНАЯ ЛИНГВИСТИКА И ОБЩАЯ ТЕОРИИ ЯЗЫКА

В статье обсуждаются направления развития корпусной лингвистики и её влияние на теорию языка. Автор выделяет три основных подхода: анализ, использующий корпус, основанный на корпусе и направляемый корпусом, и рассматривает их методологические и практические различия. Автор обсуждает, что корпусная лингвистика может дать общей теории языка и какие новые вопросы ставит перед лингвистами.

Ключевые слова: корпусная лингвистика, идиоматичность, количественный анализ, дистрибутивная семантика, конкурирующая мотивация, антропоцентричный подход, большие языкове модели.

В современном употреблении можно выделить два основных значения термина «корпусная лингвистика»: 1) создание корпусов; 2) анализ языкового материала с помощью корпуса. В настоящем докладе я буду говорить только о корпусной лингвистике во втором значении.

Можно выделить три подхода к корпусным исследованиям.

— Анализ, использующий корпус [англ. corpus-informed analysis] предполагает использование корпусов для сбора примеров естественного языка без количественного анализа. Этот метод снижает роль интуиции исследователя и позволяет более объективно анализировать языковые данные, найденные в корпусе.
— Анализ, основанный на корпусе [англ. corpus-based analysis] комбинирует качественные и количественные методы для проверки лингвистических гипотез. Этот подход предполагает хотя бы минимальный количественный анализ, а в идеале — статистический анализ и воспроизводимость результатов.
— Анализ, направляемый корпусом [англ. corpus-driven analysis] исключает (полностью или частично) изначальные теоретические допущения, фокусируясь на автоматизированной обработке

данных. Именно этот подход привел к созданию «больших языковых моделей».

За десятилетия своего существования корпусная лингвистика накопила большое количество наблюдений, которые подтверждают, уточняют и даже опровергают некоторые представления о языке. Далее я перечислю те идеи, которые, на мой взгляд, внесли наиболее значительный вклад в общую теорию языка.

1. Представление об иерархической структуре языка часто сталкивается с фактами, которые не вписываются в строгую линейную систему иерархий. Эти факты обычно рассматривались как исключения из правил и маргинализовались в области фразеологии. Однако корпусная лингвистика и модели языка, основанные на употреблении, показали, что идиоматизация играет гораздо более заметную роль, размывая границы между языковыми уровнями [Sinclair 1991].

2. К настоящему моменту имеется множество доказательств того, что наша речь не порождается каждый раз заново, а в значительной степени состоит из (полу)готовых фраз, которые мы храним в памяти. Длинные синтаксические связи, вероятно, требуют грамматической поддержки на уровне правил, однако на уровне малого синтаксиса сочетаемость единиц оказывается устойчивой и предсказуемой без обращения к правилам [Hunston 2000].

3. Линейное развертывание речи играет важную роль как в производстве, так и в восприятии языка. Это означает, что в потоке речи каждый произнесенный/написанный языковой знак снижает неопределенность, тем самым увеличивая предсказуемость высказывания. Синтагматические связи, таким образом, играют гораздо более фундаментальную роль, чем это предполагалось ранее [Sinclair, Mauranen 2007].

4. Для обобщения новых наблюдений было предложено объяснение, согласно которому при формировании высказывания говорящий не следует последовательно от глубинных структур к их поверхностной реализации. Вместо этого он выбирает между несколькими вариантами построения высказывания, не разделяя их на языковые уровни, а отбирая наиболее подходящие элементы из любого доступного языкового материала. Эта процедура получила название «конкурирующей мотивации» [MacWhinney 2014].

5. Таким образом, в дихотомии «язык-речь» роль синтагматических связей значительно возросла. Речевая деятельность понимается не просто как основная по отношению к языковой системе — стирается и сама граница между правилами и их реализацией, между языком и речью. Наиболее адекватным подходом к описанию языка оказывается не моделирование взаимодействия языковых единиц, разделённых на уровни, а описание всех — как индивидуальных, так и самых общих — параметров употребления. Эти параметры представляют собой единый континуум, в котором разграничение между языком и речью становится условным [Goldberg 2006].

6. Лингвистическая теория, основанная на правилах, не смогла достичь впечатляющих результатов в создании работающих приложений, а успех больших языковых моделей поддерживает скорее вероятностный подход к порождению высказывания. Если во второй половине 20-го века инженеры опирались на достижения лингвистов; то сейчас лингвисты должны, если не опираться на достижения инженеров, то как минимум объяснить, почему большие языковые модели работают так хорошо без лингвистической теории [Sutton 2019].

7. Антропоцентричный подход к описанию языка, без сомнений, имеет право на существование. Однако он имеет тенденцию усложнять анализ так, что становится менее пригодными для создания универсальных решений. Горький урок, преподнесенный большими языковыми моделями, состоит в том, что возможно построение эффективных систем, основанных на альтернативных способах порождения высказывания. Считать ли ChatGPT «банальностью зла, которая просто следует приказам» (Chomsky 2023) или новым объектом лингвистического анализа — решать нам.

Литература

Chomsky, N. The false promise of ChatGPT. // The New York Times. 2023. № 8. [Электронный ресурс]. URL: https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Goldberg, A. Constructions at Work: The Nature of Generalization in Language. London: Oxford University Press. 2006.

Hunston, S. Grammar: A Corpus-driven Approach to the Lexical Grammar of English. Amsterdam: John Benjamins Publishing. 2000.

MacWhinney, B. Competing Motivations in Grammar and Usage. London: Oxford University Press. 2014.

Sinclair, J. Corpus, Concordance, Collocation. Oxford: Oxford University Press. 1991.

Sinclair, J. M., Mauranen, A. Linear Unit Grammar: Integrating Speech and Writing. London: John Benjamins Publishing. 2007.

Sutton, R. The Bitter Lesson. // Incomplete Ideas. 2019. [Электронный ресурс]. URL: http://www.incompleteideas.net/IncIdeas/BitterLesson.html

К. М. Корчагин
(Москва, Россия)
Институт русского языка им. В.В. Виноградова РАН
stivendedal@gmail.com

ПОЭТИЧЕСКИЙ КОРПУС НКРЯ ДВЕ ДЕКАДЫ СПУСТЯ: РЕЗУЛЬТАТЫ И ПЕРСПЕКТИВЫ

Поэтический подкорпус НКРЯ существует чуть меньше основного корпуса и с самого начала задумывался как инструмент для исследователей русской поэзии и поэтического языка, то есть должен был решать более локальную задачу, чем основной корпус. На сегодняшний момент поэтический подкорпус содержит обширную коллекцию русской поэзии XVIII—XXI веков, отражает все заметные поэтические направления и продолжает постоянно пополняться. Этот корпус примечателен своей специфической разметкой — стиховедческой, отражающей ключевые параметры поэтического текста, среди которых метр, строфика, схема рифмовки и другие. Эти параметры позволяют производить поиск и структурировать массивы русских поэтических текстов с совпадающими формальными параметрами, автоматическим образом выполняя задачу, для решения которой раньше требовались специализированная справочная литература и/или трудоемкая обработка текстов. А использование стиховедческой разметки вместе с грамматической позволяет пользователю корпуса решать задачи, лежащие на границах лингвистики и литературоведения. В докладе будет рассмотрено, какой путь прошел поэтический подкорпус за два десятилетия своего существования, какие задачи с помощью него могут быть решены уже сейчас и перспективы открываются при дальнейшем его усовершенствовании.

О.Ю. Крючкова, А.И. Буранова
(Саратов, Россия)
Саратовский государственный университет
имени Н.Г. Чернышевского
vpks@rambler.ru, aburanova@list.ru

ДИАЛЕКТНЫЙ КОРПУС КАК РЕСУРС КОММУНИКАТИВНОЙ ДИАЛЕКТОЛОГИИ

Использование текстовых диалектологических корпусов позволяет осуществлять текстоориентированный подход к исследованию диалектной речи, дает наиболее достоверный материал для выявления особенностей ее построения, определения квантитативных отношений языковых единиц в диалектной коммуникации, функционального анализа внутридиалектной и междиалектной вариативности. Корпусные исследования диалектной речи знаменуют переход от констатирующего (таксономического) описания диалекта к функциональному; от дифференциального изучения диалекта к его исследованию как самостоятельного, полноценного культурно-коммуникативного образования.

Ключевые слова: текстовый корпус, диалектная речь, коммуникативная диалектология

Бурное развитие технологии и методологии корпусных лингвистических исследований поставило на повестку дня создание диалектных текстовых корпусов, предоставляющих репрезентативный материал для многостороннего изучения специфики диалектных подсистем и специфики общения на диалекте. Диалектные корпуса такого типа должны обладать свойствами лингвистической модели, отражать в необходимых пропорциях черты моделируемого объекта. Подобная цель может быть достигнута в диалектных корпусах, создаваемых на материале отдельных говоров и представляющих говоры как самодостаточные коммуникативных системы. Моделирование коммуникации в конкретном говоре требует включения в корпус целых речевых произведений (текстов), полно отражающих особенности диалектной речи: ее важнейшие формы, жанрово-тематическую структуру, социальную и идиолектную дифференциацию (о принципах организации базы данных диалектологического корпуса см. [Крючкова, Гольдин, 2008; Крючкова, Гольдин, 2011]).

Текстовый диалектный корпус позволяет осуществить переход от констатирующего (таксономического) описания диалекта к функциональному. На основе корпусных данных исследователь может получить не только актуализованный в речи инвентарь интересующих его форм, но и надежные сведения об их функциональных особенностях: о частотности явления в целом и относительной частоте репрезентирующих это явление форм в данном говоре, в идиолектах его носителей, а при наличии целого ряда корпусов отдельных говоров — в системах разных говоров (получение таких данных даст основания для разграничения окказионального и узуального в говоре, идиолектного и общедиалектного); о функциональном соотношении вариантов, об их типичной и нетипичной сочетаемости (контекстной реализации) в данном говоре, в идиолектах его носителей, в системах разных говоров; о характере функционирования языкового элемента в разных по тематике и жанру речевых сегментах.

Так, например, функциональный анализ морфологических диалектизмов в корпусе среднерусского говора показывает незначительное морфологическое своеобразие среднерусской диалектной речи. Словоформы, не соответствующие литературной норме, составляют менее 1% от общего числа словоупотреблений (их вклад в коммуникацию на диалекте невелик), хотя в таксономическом отношении они разнообразны. Кроме того, выборка по литературным соответствиям позволяет заключить, что ни один из отмеченных типов морфологических диалектизмов не обладает в говоре абсолютной регулярностью, т.к. для каждой из диалектных словоформ отмечены совпадающие с литературной нормой варианты (*дожди-**ти** были*, но: *а так оне… родники-**то** кто знает*; *она/ хворала/ **у ней**… лёгкие больные были*, но: *и вот здесь **у неё** это… такие вот/ прям/ нарывы/ кругом*), подробнее см. [Крючкова, Гольдин, 2007]. Иную картину показывают тексты севернорусского говора, в котором диалектная морфология представлена более широко и регулярно.

Корпусный анализ лексического своеобразия диалектной речи (выборка единиц, отмеченных признаком дифференциальности по отношению к литературному стандарту) позволяет выделить типы лексической нестандартности, относительный вес каждого из типов нестандартных единиц в диалектной речи, наличие и характер

вариативности при их использовании (в том числе коммуникативное соотношение с формами литературного стандарта), см., например, [Крючкова 224].

Исследование количественного распределения грамматических классов слов и грамматических форм (напр., падежных форм имен существительных, темпоральных форм глагола) в диалектной речи дает возможность охарактеризовать диалект как идиом, отличающийся по названным признакам от литературной речи (в том числе литературно-разговорной) [Буранова 2015].

Тематическая и жанровая разметка значительного по объему корпуса диалектных текстов позволяет выявлять соотношение различных тем и жанров в составе диалектной коммуникации. Различные предметные области (*частная жизнь, дом и домашнее хозяйство, религия, зрелища и развлечения, политика и общественная жизнь, производство*) занимают в диалектной коммуникации неодинаковое место и с точки зрения объема, и с точки зрения их смысловой детализации. Выделенные предметные области различаются также особенностями вербализации соответствующей тематики, своеобразием лингвистического развертывания темы (см. подробнее [Гольдин, Крючкова, 2006]), что, в свою очередь, указывает на степень когнитивной актуальности и коммуникативной релевантности предметной области.

Реализация тематического и пословного поиска в диалектном корпусе в сочетании с данными о частотности словоформ позволяет выделить актуальные для диалектной коммуникации концептуальные переменные, а конкордансы запрашиваемых словоформ предоставляют ценный материал для содержательного анализа соответствующих концептов и концептуальных оппозиций. Ограничение поиска подкорпусом, соответствующим отдельному говору, предоставляет данные о полидиалектном vs. монодиалектном характере каждой из выделенных оппозиций.

Использование текстовых корпусов диалектной речи дает возможность перейти от дифференциального анализа диалекта к его исследованию как самостоятельного, полноценного культурно-коммуникативного образования, осуществить полнообъемное описание коммуникации на диалекте.

Литература

Буранова А.И. Количественные признаки языковых идиомов: диалектная речь на фонге литературно-разговорной (на материале русского языка) : дис. ….канд. филол. наук / Сарат. гос. ун-т, Саратов, 2015. 211 с.

Гольдин В.Е., Крючкова О.Ю. Тематическая разметка и тематический анализ диалектного текстового корпуса // Языковая личность — текст — дискурс: теоретические и прикладные аспекты исследования. — Ч.1. — Самара: изд-во «Самарский университет», 2006. С. 71 — 80.

Крючкова О.Ю. Некодифицированные лексические элементы в речи диалектоносителей // Труды Института русского языка им. В.В. Виноградова. 2024. № 3(41). С. 213–220.

Крючкова О.Ю., Гольдин В.Е. Текстовый диалектологический корпус как модель традиционной сельской коммуникации // Компьютерная лингвистика и интеллектуальные технологии. — Вып. 7 (14). М.: РГГУ, 2008. С. 268-273.

Крючкова О.Ю., Гольдин В.Е. Корпус русской диалектной речи: концепция и параметры оценки // Компьютерная лингвистика и интеллектуальные технологии. Вып. 10 (17). М.: РГГУ, 2011. С. 359-367.

О.Ю. Крючкова

(Саратов, Россия)
Саратовский государственный университет
имени Н.Г. Чернышевского
vpks@rambler.ru

ВКЛАД ВАЛЕНТИНА ЕВСЕЕВИЧА ГОЛЬДИНА В РАЗВИТИЕ КОРПУСНОЙ ДИАЛЕКТОЛОГИИ

Концепция создания и принципов построения корпуса диалектных текстов начинает развиваться В.Е. Гольдиным с конца 1980-х годов, в период обсуждения структуры Машинного фонда русского языка. Теоретической базой корпусной концепции В.Е. Гольдина стала разработка идей коммуникативной диалектологии, основным объектом которой становятся говоры как самостоятельные, целостные коммуникативные системы, как особый тип речевой культуры. Главным объектом наблюдений становятся связная речь и воплощенные в ней единицы общения, тексты, а основным источником материала — корпуса отдельных говоров.

Ключевые слова: текстовый корпус, диалектология, В.Е. Гольдин.

Вопрос о создании диалектных корпусов был поставлен в конце 1980-х гг., в период обсуждения концепции создания Машинного фонда русского языка (МФРЯ) [Машинный фонд… 1986]. А.С. Гердом и В.Е. Гольдиным были высказаны идеи о создании диалектологического подфонда МФРЯ: об ориентации данного подфонда на полные магнитофонные записи диалектных текстов (А.С. Герд); о текстовых корпусах отдельных говоров как основе диалектологического подфонда МФРЯ (В.Е. Гольдин). В сборнике материалов III Всесоюзной конференции по созданию машинного фонда русского языка была опубликована статья В.Е. Гольдина «К проекту текстового диалектологического подфонда Машинного фонда русского языка» [Гольдин 1990].

Концепция машинного текстового диалектологического фонда базировалась на разрабатываемой В.Е. Гольдиным теории коммуникативной диалектологии. Накопленный в 1980-е годы массив аудиозаписей диалектной речи, значительный объем их текстовых расшифровок способствовали пониманию того, что русские народные

говоры целесообразно изучать не только в лингвогеографическом аспекте, но и как самостоятельные коммуникативные системы, представляющие особый тип речевой культуры. Наметилось смещение интереса диалектологов с междиалектных различий на то, что характеризует общую специфику «диалектов как диалектов». В работах В.Е. Гольдина стало оформляться особое направление — «коммуникативная диалектология», и тексты стали осознаваться как основной информационный ресурс этого научного направления. Положения коммуникативной диалектологии и утверждение в качестве ее основных источников текстовых диалектных корпусов были обоснованы В.Е. Гольдиным в работах «Диалектологический текстовый машинный фонд говора и исследование диалектных изменений» [Гольдин 1991], «Машиннообрабатываемые корпусы диалектных текстов и проблема типологии русской речи» [Гольдин 1995], в его докторской диссертации «Теоретические проблемы коммуникативной диалектологии» [Гольдин 1997].

Коммуникативная диалектология направляет внимание на новые для диалектологии объекты исследования. К ним относятся:

– информационная структура общения на диалекте (например, характерный для диалектной речи набор речевых событий, речевых жанров, специфика реализации универсальных (общих для национальной культуры) речевых событий и жанров, событийные и жанровые «лакуны»);

– особенности текстовой деятельности на диалекте (особенности построения текстов, связанные во многом с социальной организацией диалектного коллектива);

– когнитивная сторона общения на диалекте (особенности картины мира носителей традиционной деревенской культуры, характер выражаемых в речи знаний, способы их речевого представления);

– место речи в составе деятельности носителей традиционной культуры и характер рефлексии диалектоносителей над речью;

– особенности сохранения речевой традиции, специфика трансляции ее во времени, характер прецедентных текстов диалектного общения.

В. Е. Гольдин по-новому определил сущность диалекта как особого языкового образования. Она заключается прежде всего не в

отличиях одних говоров от других, а именно в признаках, объединяющих диалектную речь любых территорий бытования данного языка и при этом характерных не для всех языковых стратов, — в признаках, отличающих диалектную речь от литературной в первую очередь. Выделение и описание этих признаков и составляет специфический объект коммуникативной диалектологии.

В. Е. Гольдин подчеркивал, что коммуникативное изучение диалектной речи (решение сформулированных им задач коммуникативной диалектологии) требует специальной организации научных источников: главным объектом наблюдений становятся связная речь и воплощенные в ней единицы общения, тексты. Таким образом, коммуникативная диалектология ведет к диалектологии корпусной.

В.Е. Гольдиным выделены принципиальные положения общей архитектуры диалектного корпуса, целью которого является представление диалекта как целостного культурно- коммуникативного образования:

– корпус представляет реально функционирующие коммуникативные образования — говоры как диалектные микросистемы.

– текстовая база каждого говора-подкорпуса строится как модель диалектной коммуникации, формируется с установкой отразить важнейшие типы диалектной речи (речь бытовую, фольклорную, речь в условиях официального, обрядового общения); различные формы речи (диалог, полилог, монолог); разнообразную тематику сельского общения; социальную дифференциацию носителей говора (по полу, возрасту, профессии, уровню образования).

– диалектный корпус является корпусом лингвокультурологической направленности, в котором воссоздается лингвистический и культурный фон традиционной культуры в целом и в связи с содержанием каждого конкретного текста в отдельности.

Эта концепция получила дальнейшее развитие при создании диалектного корпуса в Саратовском государственном университете.

Литература

Гольдин В. Е. К проекту текстового диалектологического подфонда Машинного фонда русского языка // Материалы III Всесоюзной конференции по созданию машинного фонда русского языка. М.: Изд-во Моск. ун-та, 1990. С. 92–103.

Гольдин В. Е. Диалектологический текстовый машинный фонд говора и исследование диалектных изменений // Современные процессы в русских народных говорах. Саратов: Изд-во Сарат. ун-та, 1991. С. 17–28.

Гольдин В.Е. Машиннообрабатываемые корпусы диалектных текстов и проблема типологии русской речи // Русистика сегодня. 1995. №3. С. 72–87.

Гольдин В.Е. Теоретические проблемы коммуникативной диалектологии : дис. в виде науч. докл. …д-ра филол. наук. Саратов, 1997. 52 с.

Машинный фонд русского языка: идеи и суждения. М.: Наука, 1986. 240 с.

В. П. Лелик[1], М.Д. Дьячкова*[1], А. С. Сычева[1], С. В. Дорофеева[1], И. А. Секерина[2]

[1] Москва, Россия, Центр языка и мозга НИУ ВШЭ;
[2] Нью-Йорк, США, Городской университет Нью-Йорка
**masha.dyachcova@yandex.ru*

КОРПУС ДЕТСКОЙ РЕЧИ В ФОРМАТЕ CHILDES: ОПЫТ СОЗДАНИЯ БАЗЫ ДАННЫХ И ПРИМЕНЕНИЯ ИНСТРУМЕНТОВ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

В докладе представлен опыт создания и использования для исследовательских целей русскоязычного корпуса детской речи в формате CHILDES. Этапы создания корпуса включают сбор данных, расшифровку и анонимизацию записей, а также автоматическую морфологическую разметку. Исследования на основе собранных корпусов охватывают различные аспекты усвоения языка, от морфологии до синтаксиса, и демонстрируют значимость CHILDES как инструмента для лингвистических исследований. В дальнейшем планируется расширение участников и углубление разметки, что открывает новые перспективы для анализа речевого развития детей.

Ключевые слова: корпус детской речи, усвоение языка, CHILDES, лонгитюдное исследование, ранние этапы усвоения языка

В 2016 году в Центре языка и мозга НИУ ВШЭ был начат сбор данных русскоязычных детей в соответствии с протоколом CHILDES. CHILDES — база данных записей устного спонтанного дискурса, основными участниками которого являются монолингвальные дети в возрасте от 1 года до 3 лет, а также их родственники [MacWhinney: 2000]. На данный момент момент к русскоязычному проекту присоединились 14 семей.

Работа с данными CHILDES включает в себя несколько этапов. Прежде всего, родители записывают свое обычное общение со своими детьми, эти записи производятся примерно один раз в две недели и длятся около 30 минут. Затем записи расшифровываются вручную в программе CLAN [MacWhinney 2000], программа позволяет соотнести каждую реплику с временным отрезком, когда она была произнесена. При расшифровке транскрибируется не только речь ребенка, но и реплики взрослых. Следующий этап работы — анонимизация транскриптов для их последующего размещения в открытом доступе, по примеру корпуса BiRCh [Lưu и др. 2022], где личные данные

участников заменяются псевдонимами, а конфиденциальная информация скрывается. Для этого, а также для анонимизации лиц на видео, была разработана программа на языке Python [Python Software Foundation, URL: https://www.python.org]. Затем проводится морфологическая разметка с использованием библиотеки MyStem. Программа автоматически определяет лемму, часть речи и грамматические признаки каждого слова. Впоследствии результаты анализа проверяются лингвистами вручную. Итоговые данные сохраняются в Excel для последующей обработки и анализа в формате Pandas DataFrame.

Для двух корпусов, собранных в Центре Языка и Мозга, уже завершены сбор данных, расшифровка и морфологическая разметка подкорпусов. Эти подкорпуса основаны на записях девочки Тоси и мальчика Яши. Записи Тоси велись с 10 месяцев до 4 лет, всего подкорпус состоит из 254 транскриптов и примерно 30 часов видео. Записи Яши велись с одного года 4 месяцев до 3 лет, подкорпус насчитывает 43 транскрипта и 12 часов видео. Остальные корпуса находятся в разной степени готовности. На данный момент корпуса Тоси и Яши готовятся к публикации в открытый доступ для широкого круга исследователей. В целом, проект CHILDES находится на этапе пополнения новыми материалами и расширения количества корпусов.

На материале уже размеченных записей речи были проведены исследования различных аспектов усвоения языка: от морфологии до синтаксиса. Например, в 2020 году мы изучали усвоение грамматических категорий глагола, в результате чего выяснили порядок их усвоения и сравнили с данными предыдущих исследований [Lelik и др. 2020: 49-50]. Далее мы обратились к изучению лексики и исследовали формирование словарного запаса [Лелик, Лопухина 2021: 526–531]. Также мы разработали Индекс продуктивности синтаксиса при освоении именных групп русскоязычными детьми и протестировали его на материале корпусных данных [Дьячкова и др. 2023]. данные корпуса позволили впервые для русского языка протестировать метрику MLU (mean length of utterance) на материале лонгитюдных данных, было обнаружено, что эта метрика успешно предсказывает речевое развитие русскоязычных детей [Дьячкова 2024]. Наконец, мы начали работу с билингвальными данными и в 2022 году провели исследование билингвального инпута [Лелик 2024].

Перспективы развития корпуса включают дальнейшее расширение числа участников, проведение разметки на других уровнях языка (фонология, синтаксис, лексика). Корпус детской речи в формате CHILDES представляет собой важный инструмент для лингвистических исследований, который имеет огромный потенциал для применения в будущем.

Литература

Дьячкова М. Д., Секерина И. А., Дорофеева С. В. Разработка и апробация Индекса продуктивности синтаксиса при освоении именных групп русскоязычными детьми // Когнитивная наука в Москве: новые исследования. 2023. С. 583–587.

Дьячкова М. Д. Оценка речевого развития русскоязычных детей до трех лет методами MLU и IPSyn на материале лонгитюдных данных // Выпускная квалификационная работа. 2024

Лелик В. П., Лопухина А. А. Ранние этапы формирования словарного запаса у русскоязычного ребенка 1–3 лет (на материале корпуса CHILDES) // Когнитивная наука в Москве: новые исследования. 2021. С. 526–531.

Лелик В. П. Роль языкового инпута на ранних этапах развития речи детей-билингвов // Выпускная квалификационная работа. 2024

Lelik V., Lopukhina A., Korkina I. Early Stages of the Acquisition of Verbal Grammar by Russian-speaking 1-to-3-year-old Children (Based on the CHILDES Corpus) // Proceedings of the International Conference on Language Acquisition. 2023. P. 49–50.

Luu A., Koval P., Malamud S. A., Dubinina I. Y. Creating a large-scale audio-aligned parsed corpus of Bilingual Russian Child and Child-Directed Speech (BiRCh): Challenges, solutions, and implications for research // Bakhtiniana: Revista de Estudos do Discurso. 2022. Vol. 17, № 4. P. 223–261.

MacWhinney B. The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah: Lawrence Erlbaum Associates, 2000.

Python Software Foundation [Электронный ресурс]. URL: https://www.python.org/

А.Б. Летучий

(Москва, Россия)
НИУ ВШЭ
Институт русского языка им. В.В. Виноградова РАН
alexander.letuchiy@gmail.com

РУССКИЙ/Е ПАССИВ(Ы) С ПРИЧАСТИЕМ: КАКИЕ ОНИ БЫВАЮТ И КАК ИХ ИССЛЕДОВАТЬ?

В русском языке есть два типа способов выражения пассива. Один из них — конструкции, содержащие пассивное причастие (как полное пассивное причастие — *Я увидел* ***разрушенное*** *здание* — так и конструкции с кратким причастием и с глаголом *быть*, либо без глагола — *Здание уже давно* ***(было) разрушено***). Второй — пассивное прочтение возвратных форм (*Дом ещё* ***строится***).

С каждым из этих способов связаны свои проблемы. Возвратные формы представляют собой проблему из-за развитой полисемии — не всегда очевидно, можно ли причислять к пассивам образования типа *Здесь* ***продаётся*** *мясо* или *Улица* ***называется*** *Тверская*. Однако мы сконцентрируемся на вопросах, связанных с пассивным причастием и конструкциями, в которых оно участвует.

Стандартной пассивной конструкцией с глаголом *быть* обычно считаются именно конструкции типа *здание разрушено* с кратким причастием:

(1) *На узловую станцию Лиски — последнюю, где состав был замечен, — срочно прибыл наряд ОГПУ. [Гузель Яхина. Дети мои (2018)]*

Конструкции с полным причастием вида *чашка была разбитой* нередко считают скорее не единым способом выражения пассива, а сочетанием, где пассив выражает формой *разбитой*, а *быть* употребляется в стандартном значении прошедшего времени. Во многом этот вывод подтверждается семантикой — так, в примере (2) выражается стативное значение, оно кодируется причастием *испорченной*, а глагол *быть* просто относит это состояние к плану прошедшего.

(2) *Да и пища часто была* ***испорченной****: хлеб плесневел, масло прогоркало, в мясе и рыбе заводились черви...* [С. Матссон-Попова. Корабль «Васа» // «Наука и жизнь», 2007]

Однако проблемы возникают в том случае, когда синтаксический контекст либо не позволяет употребить конструкции с *быть* + краткими причастиями, либо их употребление сомнительно. К таким, в частности, относятся конструкции, где *быть* выступает в составе инфинитивного оборота и подчинён матричному глаголу типа *хотеть* или *бояться*. В этом случае важно обращение к корпусу — корпусные данные показывают, что при разных матричных глаголах пассивная форма ведёт себя по-разному.

Таблица 1. Распределение конструкций типа *должен* + *быть* + полное причастие и *должен* + *быть* + краткое причастие при разных матричных предикатах (основной корпус НКРЯ, примеры с 1950 года).

Глагол	**Полное причастие**	**Краткое причастие**
бояться	83	0
должен	73	10 604
хотеть	147	29
мочь	203	18 891
надеяться	9	0
пытаться	1	0
стараться	6	0
желать	44	2

Оказывается, что при предикатах *должен* и *мочь* всё равно предпочитается сочетание *быть* с краткой формой; при *хотеть* краткая форма встречается реже полной, однако всего в 5 раз; при остальных предикатах краткая форма не встречается или почти не встречается. Эти различия допускают два объяснения — не конфликтующие, а связанные между собой. Во-первых, краткую форму предпочитают или хотя бы с ней сочетаются предикаты, в наибольшей мере похожие на вспомогательные глаголы (например, *мочь* и *должен* могут выражать разные типы модальности и не ограничивают одушевлённость и семантический тип субъекта). Во-вторых, краткая форма встречается тем чаще, чем выше частотность пассивной конструкции с данным предикатом в целом: *должен* (всего 10677 примеров) и *мочь* (19 094) существенно обгоняют по частотности все остальные предикаты, далее следует *хотеть* (176), для остальных предикатов совокупная встречаемость конструкций с полным и кратким причастием составляет менее сотни примеров.

При этом у тех предикатов, для которых встречается почти исключительно полная форма, она имеет значение, не характерное для неё при финитных формах *быть* — значение динамической завершённой ситуации (3):

(3) — *Предупреждаю, — кричал он мне, пытаясь* ***быть услышанным*** *на фоне оглушающего музыкального ритма, — ты окажешься в настоящем загоне для скота.* [Посланники дьявола (2004) // «Солдат удачи», 09.06.2004] (ср. сомнительное *Он был услышанным,* гораздо лучше *Он был услышан*).

Напротив, при *должен* и *мочь* динамическое значение имеет конструкция с кратким причастием (4), а конструкция с полным (5) имеет обычно стативное значение:

(4) *<…> которому также должен* ***быть сообщен*** *PIN-code.* [Денежные переводы мигрантов -- фактор инновационного развития мировой финансовой инфраструктуры // «Вопросы статистики», 2004]

(5) *А я должен* ***быть одетым*** *как ты, когда убиваю Гермеса.* [Алексей Иванов. Комьюнити (2012)]

Можно сделать вывод, что по крайней мере в части синтаксических контекстов (с глаголами типа *бояться, желать* и, видимо, *хотеть*) конструкция с полной формой является синтаксической трансформацией финитной конструкции с краткой формой (*Он был услышан — Он пытался быть услышанным*) и выражает пассив как единое целое. В других контекстах, например, *Я должен быть одетым как ты* следует считать, что пассив выражается только формой *одетый* — а конструкция со вспомогательным глаголом выглядит как *должен быть сообщен* (см. пример (4) выше). При этом функционирование конструкций с полной и краткой формой зависит от статуса и частотности матричного глагола.

В докладе мы обсудим и другие статистические и грамматические свойства пассивных конструкций, которые помогает найти обращение к корпусу.

О. Н. Ляшевская
(Москва, Россия)
НИУ ВШЭ, ИРЯ РАН им. В. В. Виноградова
olesar@yandex.ru
С. А. Ребриков
(Москва, Россия)
НИЦ "Курчатовский институт", НИУ ВШЭ

ЛЕММАТИЗАЦИЯ ИЛИ ДИЗАМБИГУАЦИЯ? К ПРОБЛЕМЕ ЛЕКСИКО-ГРАММАТИЧЕСКОГО АНАЛИЗА СОКРАЩЕНИЙ В НКРЯ

В докладе рассматривается проблематика автоматической обработки словоформ, являющихся инициальными аббревиатурами и сокращениями с точкой. Не имея явно выраженных показателей словоизменения, эти единицы представляют сложный случай для моделей лемматизации и морфологического и синтаксического анализа, несмотря на то, что контекст, как правило, позволяет легко восстановить их лексико-грамматические характеристики. Мы представляем набор данных и результаты экспериментов, связанных с нейросетевым моделированием разбора сокращений в текстах корпуса.

Ключевые слова: лемматизация, дизамбигуация, лексико-грамматический анализ, автоматическая обработка текста, НКРЯ

Несклоняемые инициальные типы аббревиатур (ТК, СПбГУ) и сокращения вида *т.* (где *т.* может означать *телефон*, *то, так, тетя, товарищ, твой* и т. д.) отличаются от склоняемых аббревиатур (ср. *МИД, МИДа, МИДу, дисбат, дисбата*) отсутствием выраженных на письме окончаний. С точки зрения лексико-грамматического разбора в корпусе это означает, что словоформе *ТК* или *т.* может быть потенциально приписано более десятка морфологических разборов (например, представляющие разные комбинации числа и падежа) — явно большее, чем в среднем у обычной, склоняемой словоформы. Морфологическая омонимия может накладываться на лексическую омонимию, характерную для сокращений: «число возможных комбинаций между символами алфавита на несколько порядков меньше числа комбинаций между словоформами определенного языка — таким образом, процесс аббревиации чисто комбинаторно повышает

вероятность возникновения неоднозначности» [Гусяцкая 2024]. Кроме того, сокращения могут быть определены как более динамическая часть лексикона [Zahariev 2004], а это значит, что далеко не все сокращения можно найти в существующих словарях и в принципе учесть в словарной поддержке системы автоматической обработки текстов корпуса.

Сокращения «с точкой» (которые в корпусе могут быть записаны и без точки — например, в текстах социальных сетей) проблематичны и еще в одном аспекте. В стандарте Основного корпуса со снятой вручную омонимией и ряда других корпусов принято расшифровывать лемму до полного вида. Тем самым, при обучении нейросетевых моделей лемматизации в обучение попадают пары вида *г — город*, *г — год*, и система «видит» псевдоокончания вида *-ород*, *-од*. В результате, могут быть построены ошибочные разборы лемм вида *хород* для сокращения *х.* или *челадемик* для сокращения *чел.* в системах, построенных как на классификации, так и на генерации лемм [Lyashevskaya et al. 2024], [Morozov et al. ms.].

Таким образом, велика вероятность того, что сокращения рассматриваемых типов будут разобраны неправильно в корпусе при автоматическом анализе или будут источником «странного» контекста, ведущего к неправильному анализу соседних слов. В этой связи можно вспомнить практику приписывания сокращениям отдельной части речи «аббревиатура», грамматической характеристики «аббревиатура» (вместо характеристик рода, числа, времени и т.п.), леммы, равной сокращению, существующую в ряде корпусов мира и позволяющую так или иначе уйти от проблемы анализа сокращений.

В докладе будет представлен набор данных на основе материалов НКРЯ для анализа сокращений с реконструкцией полной леммы и полного грамматического разбора по контексту. Будут проанализированы сильные и слабые стороны автоматических методов лексико-грамматического анализа, основанных на словарном принципе, на обучении с учителем в архитектурах семейства BERT и BART, а также на подходе few-shot learning больших языковых моделей.

Литература

Гусяцкая П. А. Автоматическое разрешение неоднозначности аббревиатур в русскоязычном корпусе медицинских текстов. Выпускная квалификационная работа. СПб.: СПбГУ, 2024.

Lyashevskaya, O., Afanasev, I., Rebrikov, S., Shishkina, Y., Suleymanova, E., Trofinov, I., and Vlasova, N. Disambiguation in context in the Russian National Corpus: 20 years later // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции “Диалог 2023”. М., 2023.

Morozov D., Glazkova A., Afanasev I., Lyashevskaya O., Smal I., Vlasova N. Rubic2: Ensemble Model for Russian Lemmatization (submitted).

Zahariev M. A (acronyms). Doctoral dissertation. Simon Fraser University, 2004.

Е.В. Маринова
(Нижний Новгород, Россия)
НГЛУ им. Н.А. Добролюбова
marinova@list.ru

ВАРЬИРОВАНИЕ ОБОЗНАЧЕНИЙ ВЫСОКОТЕХНОЛОГИЧНЫХ ЧЕЛОВЕКОПОДОБНЫХ ПРОГРАММ В ОТНОШЕНИИ ОДУШЕВЛЁННОСТИ/НЕОДУШЕВЛЁННОСТИ: УСЛОВИЯ И ПРИЧИНЫ

Рассматриваются обнаруженные на материале «Национального корпуса русского языка» варианты в выборе формы винительного падежа единственного и множественного числа терминов сферы высоких технологий (*цифровой ассистент, бот, чат-бот, голосовой помощник, цифровой двойник* и под.). Отмечаются условия для данного грамматического варьирования, его специфика и лингвокогнитивные причины.

Ключевые слова: грамматические варианты, термины, сфера цифровых технологий, синтаксическая одушевлённость, причины варьирования, НКРЯ.

1. В процессах номинации новых объектов и реалий современных технологий последовательно отражается приписывание техническим устройствам и программам свойств живого существа (антропоморфизм). Начиная с 90-х гг. XX в., реалии техномира регулярно именуются из ресурсов **личных** существительных, прежде всего агентивов. На русской почве такие номинации строятся нередко с помощью атрибутов, имеющих значение ‘осуществляемый, существующий в интернет-пространстве’: *онлайн-переводчик* (программа), *интернет-брокер, онлайн-брокер, электронный брокер* (пользовательская программа для проведения брокерских операций в режиме реального времени через интернет), *виртуальный учитель* (обучающая программа в Сети), *цифровой актёр* (о цифровой копии актёра), *цифровой полицейский* (система безопасности, основанная на использовании ИИ, дронов, видеокамер и др.). В последнее десятилетие чрезвычайно актуальным становится в позиции перед личным существительным формант *ИИ-* ‘связанный с использованием технологии искусственного интеллекта’: *ИИ-врач, ИИ-доктор, ИИ-*

кардиолог, ИИ-терапевт, ИИ-дизайнер, ИИ-микробиолог, ИИ-композитор, ИИ-юрист и др. В этом же ряду модные в настоящее время номинации с *робо-*: *робо-коуч, робо-инспектор.*

2. Перечень этих **антропоморфных** метафор можно продолжить интернациональными терминами IT и ИИ-сферы: *компилятор* ‘программа перевода с одного машинного языка на другой’, *корректор* (*правописания*), *мастер* (*подсказок*), *навигатор,* (*текстовый*) *редактор, виртуальный ассистент/помощник, цифровой ассистент, голосовой помощник, цифровой двойник* и др. Подобные названия своеобразно отражают идею **двойственности** современного мира, в котором меняются / смешиваются роли человека и машины, усиливаются «смысловые смещения в массовом восприятии, понимании, интерпретации фундаментальной оппозиции» [Волков 2020: 747].

В этом отношении показательна «реакция» русской грамматики на регулярное использование **одушевлённых существительных для обозначения неживых объектов**. В большинстве случаев категория одушевлённости у этих номинаций утрачивается (*включить навигатор*, а **не включить навигатора*). Однако на некоторых участках систему как будто бы «сбоит», и она «выдаёт» варианты использования слова в форме винительного падежа и как одушевлённого, и как неодушевлённого. Ср.:

a. *Мы рассматриваем возможность внедрения такой функциональности в наш голосовой* ***ассистент*** *«Ева»* (Ведомости. 15.12.2021) [НКРЯ]; *«Вымпелком» тестирует виртуального* ***ассистента*** (заголовок) (Ведомости. 20.01.2021) [НКРЯ];

b. *…аппарат позволяет создать цифровые* ***двойники*** *статичных нетвердых объектов* (РИА Новости. 09.10.2020) [НКРЯ]; *Российские разработчики показали цифровых* ***двойников*** *на Венецианском биеннале* (Известия. 12.05.2019) [НКРЯ];

c. *…на смартфонах должно быть 15 приложений, включая браузер, навигатор, мессенджер, голосовой* ***помощник****, новостной агрегатор* (Парламентская газета. 30.11.2021) [НКРЯ]; *Пауль и Антон разработали голосового* ***помощника****, который не только умеет подстраиваться под характер владельца, но и незаметно делает из него идеального потребителя* (анонс фильма в приложении «Яндекс Афиша»).

Отмеченная у этих и подобных слов вариантность отличается от вариантности типа *ловил анчоусов* (о рыбе) — *ел анчоусы* (о блюде), проявляющейся в пределах семантической структуры слова между его лексико-семантическими вариантами (ЛСВ) [Крысин 1997: 69], поскольку колебания в отношении одушевлённости/неодушевлённости происходят в пределах одного и того же ЛСВ.

3. Грамматическая вариантность такого рода наблюдается сейчас и у слов *чат-бот* и в особенности *бот*, которые используются то как одушевлённые, подобно слову *робот*, то как неодушевлённые. По данным опроса установить однозначно предпочтительный вариант не представляется возможным: во-первых, результаты различны даже по формам единственного и множественного числа; во-вторых, неодинаковую частотность проявляют варианты каждого типа при разных глаголах. В этом отношении парсинг-исследование по открытым источникам показал более стройную картину: у слова *бот* преобладают формы **одушевлённого** существительного, у *чат-бот* — **неодушевлённого** (хотя преимущество последних в форме множественного числа незначительное — 51%) [Маринова 2024: 234].

Предположение о том, что рассматриваемые слова занимают место «в немногочисленной группе слов типа *вирус, призрак* или *эмбрион*» [Тельпов 2024: 106], представляется мало убедительным. Причина варьирования номинаций *бот* и *чат-бот* видится в том, что обозначаемые ими референты, с одной стороны, осмысливаются говорящими как неодушевлённые объекты (программы), с другой — персонифицируются и встраиваются в систему языка, изменяясь по образцу одушевлённого существительного *робот* (английское *bot* представляет собой усечение от *robot*). Действие таких противоположных факторов, как внеязыковая действительность и языковая система, приводит к варьированию этих номинаций.

Литература

Волков В.В. Искусственный «интеллект» и человеческий ум: футуристическая синекдоха и реальность (лингвистический и лингвоментальный аспекты) // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. 2020. Т.11. №4. С. 745–759. DOI: 10.22363/2313-2299-2020-11-4-760-774

Крысин Л.П. Лингвистическое представление иноязычного слова: типы грамматической информации // Облик слова. Сб. статей памяти Д. Н. Шмелёва. М.: РАН. Ин-т рус. яз. 1997. С. 65–71.

Маринова Е.В. Русская терминология цифрового общества: грамматические особенности в фокусе неологии и неографии // Русистика. 2024. Т. 22. №2. С. 225–239. http://doi.org/10.22363/2618-8163-2024-22-2-225-239

НКРЯ: Национальный корпус русского языка. URL: www.ruscorpora.ru

Тельпов Р.Е. Об одушевлённости слова *чат-бот* // Магия ИННО: перспективы развития лингвистики и лингводидактики в современных условиях. Т. I (№1). М.: изд-во «МГИМО-Университет», 2024. С. 106–111. DOI: 10.24833/2949-6357.2024.GEO.1

Е.В. Маркасова
(Пекин, КНР)
Пекинский университет
markasovaelena@yandex.ru

«МОЁ Я» В РУССКОМ ЯЗЫКЕ XIX-XX ВВ. ПО ДАННЫМ НКРЯ

В процессе формирования понятия «личность» в первой трети XIX века в русском языке складываются новые грамматические способы самопрезентации. Изменения в моделях самопрезентации обусловлены изменениями в психологии индивида и социума. Национальный корпус позволяет увидеть, как меняется состав предикатов при местоимении Я, как появляются тавтологии с этим местоимением, как обновляется частеречный состав риторических фигур, строительным элементом которых становится это местоимение. Этот процесс является наблюдаемым и позволяет верифицировать и детализировать некоторые сведения, полученные науками, которые не работают с такими формами доказательств, как современная корпусная лингвистика.

Ключевые слова: местоимение «я», Я-сфера, идентичность, самопрезентация, диахроническая психология, социальная психология

На протяжении последних двух веков произошли большие перемены как в восприятии человеком самости, так и в изучении психологии личности [Бергер, Лукман 1995; Порус 2012; Брюшинкин 2012; Овчинникова, Селюгина 2012; Леонтович 2017; Olson 2020]. Развитие эволюционной и диахронической психологии и возникновение в гуманитарном знании Я-проблематики требуют участия лингвистики в изучении человеческого «Я» на языковом материале. С одной стороны, в исследовании практик самопрезентации и самоидентификации лингвистика не может конкурировать с социологией и психологией. Эти науки опираются на собственные методы полевых исследований и имеют богатый опыт проведения разнообразных экспериментов, позволяющих описывать специфику поведения человека [Дудина, Смирнова 2014]. С другой стороны, у лингвистики есть корпусной подход и материал, который может быть полезен для изучения Я-проблематики.

Если мы последовательны в понимании связи между языком и действительностью, то должны признать, что изменения в жизни

общества и в психологии человека приводят к изменениям в использовании многих языковых единиц, и это утверждение банально. Однако предположение, что изменения в психологии личности влияют на функционирование местоимения «Я» в текстах, обычно вызывает критику, поскольку «связь между психологией человека и местоимением Я надо еще доказать». На наш взгляд, именно НКРЯ позволяет и доказать эту связь, и детализировать наши представления о «Я-сфере» в диахронии, и обогатить Я-проблематику анализом «Я»-сочетаемости.

С середины двадцатых годов XIX века начинается интенсивное распространение адъективных предикатов при местоимении первого лица единственного числа: «Я» постепенно освобождается от квалификатива-существительного (*Я хороший / Я хороший врач* и др.) (Маркасова 2023). Теперь прилагательное характеризует личность вне связи с ее типом деятельности или социальной сферой. Предикаты в составе конструкции «Я+адъективный предикат» характеризуют состояние Я по определенному параметру, позволяющему претендовать на включение в искомую группу или формировать группу сочувствующих. Изменения в лексическом наполнении и частоте употребления этой конструкции обнаруживают зависимость от событий, воздействующих на повседневность социума и внутренний мир человека. Наши наблюдения коррелируют с выводами специалистов в области диахронической психологии, отмечавших зависимость самосознания от внешних факторов.

К концу 1840-х гг. начинается распространение тавтологических конструкций с местоимением «я» (*я есть я, я это я, я не я*) [Лю, Маркасова 2021]. Тавтологические конструкции обычно связаны с переживанием персональной идентичности, тогда как конструкция «Я+адъективный предикат» способна отражать и процесс поиска социальной идентичности, и размышления о персональной идентичности. Использование обеих конструкций обусловлено изменениями в восприятии своей идентичности, которые возникают на фоне изменений в структурах социальных групп.

Изменения в функционировании местоимения Я происходят и на стилистическом уровне: это местоимение становится строительным материалом для риторических фигур [Маркасова, Митрофанова 2024].

Колебания количества Я-конструкций с 1800 по 2020 гг. имеют тенденцию к циклизации. Данные НКРЯ и других корпусов позволяют верифицировать выводы, полученные в дисциплинарном поле гуманитарных наук, занимающихся изучением «Я-концепции» и типов идентичности.

Литература

Olson E. T. Personal Identity // The Stanford Encyclopedia of Philosophy. 2019. Ed. N. Zalta. https://plato.stanford.edu/archives/fall2019/entries/identity-personal (дата обращения: 09.07.2020).

Бергер П., Лукман Т. Социальное конструирование реальности. М. : Медиум, 1995. 323 с.

Брюшинкин В.Н. Особенности исследования идентичности // Субъективность и идентичность. М. : Издательский дом Высшей школы экономики, 2012. С. 261–273.

Дудина В.И., Смирнова Е.Э. (ред.) Методология и методы социологического исследования. Под ред. В. И. Дудиной, Е. Э. Смирновой. СПб. : Изд-во СПбГУ, 2014. 388 с.

Иванова Н.Л. Социальная идентичность в различных социокультурных условиях // Психологический журнал. 2004. Т. 25. (1). С. 52–60.

Леонтович О.А. «Зеркало, в котором каждый показывает свой лик»: дискурсивное конструирование идентичностей // Вестник Российского университета дружбы народов. Серия: Лингвистика. 2017. Т. 21. № 2. С. 247–259.

Лю Г., Маркасова Е.В. "Я есть я (идентичность и коммуникация)" // Коммуникативные исследования. 2021. Т. 8 (4). С. 701–716.

Маркасова Е.В. «Я несчастный, но живой»: адъективный предикат при местоимении «Я» // Вестник Томского государственного университета. Филология. 2023. (82). С. 103–117.

Маркасова Е.В. Адъективный предикат «плохой/хороший» при местоимении «Я» // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. 2023. Т. 9. № 4 (36). С. 23–35.

Маркасова Е.В., Митрофанова О.А. Геминация (тройной повтор) по корпусным данным // Труды ИРЯ им. В.В.Виноградова. 2024. Т.4. С. 41–59.

Норман Б., Плотникова А. Семантика конструкций со значением социальной самоидентификации и самопрезентации в русском языке // Quaestio Rossica. 2016. Т. 4, № 4. С. 107–120.

Овчинникова, Ю. Г., Селюгина, П. Б. Личностная идентичность: от философских истоков к психологической сущности. // Психология. Журнал Высшей школы экономики. 2012. № 9 (1). С. 153–161.

Порус В.Н. (ред.) Проблема «Я»: философские традиции и современность. Под ред. В.Н. Поруса М. : Альфа-М, 2012. 351 с.

[1]Д.А. Морозов, [2]А.В. Глазкова, [3]Я.Н. Губарькова,
[4]Т.А. Гарипов, [5]С.С. Столяров, [6]Н.А. Власова,
[7]О.Н. Ляшевская, [8]И.А. Смаль, [9]А.Д. Козеренко

[1,4,5,8](Новосибирск, Россия)
НГУ, НКРЯ
[2](Тюмень, Россия)
ТюмГУ
[3,7,9](Москва, Россия)
[3]Яндекс
[7]Высшая школа экономики
[7,9]ИРЯ им. В. В. Виноградова РАН
[6](Переславль-Залесский, Россия)
ИПС им. А. К. Айламазяна РАН
[1]morozowdm@gmail.com, [2]a.v.glazkova@utmn.ru, [3]karmastina-ya@yandex-team.ru, [4]garipov154@yandex.ru, [5]s.stolyarov@g.nsu.ru, [6]nathalie.vlassova@gmail.com, [7]olesar@yandex.ru, [8]vanasmal@mail.ru, [9]akozerenko@mail.ru

ПРИМЕНЕНИЕ ИНСТРУМЕНТОВ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА НА БАЗЕ МАШИННОГО ОБУЧЕНИЯ ПРИ РАЗРАБОТКЕ КОРПУСОВ: ОПЫТ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА

В докладе представлен обзор инструментов обработки естественного языка на базе машинного обучения, используемых в Национальном корпусе русского языка. Описаны применяемые алгоритмы токенизации, лемматизации, морфологической, синтаксической, морфемной разметки, алгоритмы разметки тематики, типа, жанра и тональности текста.

Ключевые слова: машинное обучение, обработка естественного языка, Национальный корпус русского языка.

Созданный более 20 лет назад, Национальный корпус русского языка (НКРЯ) всегда использовал инструменты автоматической разметки текстов. Это связано в первую очередь с тем, что объёмы корпуса многократно превышают разумные пределы ручной разметки. На ноябрь 2024 г. общий объём текстов в НКРЯ превышает 2,2 млрд словоупотреблений, причём вручную из них размечены менее 10 млн

[Савчук и др. 2024]. С ранних этапов и до сих пор для разметки в НКРЯ используется программа MyStem [Segalovich 2003]. В то же время бурное развитие инструментов обработки языка позволило существенно развить как качество, так и разнообразие разметки Корпуса. В докладе кратко описаны применяемые в НКРЯ инструменты автоматической разметки на базе машинного обучения.

Работа с внутритекстовой разметкой всегда начинается с сегментации текста на предложения и слова. Наличие в НКРЯ текстов с ненормативной орфографией и различного рода сокращений значительно усложняет автоматическую токенизацию. Нами был проведён ряд экспериментов, в ходе которых лучшее качество продемонстрировала модель с архитектурой Stanza [Qi 2020], обученная на подготовленной выборке размеченных вручную текстов. Качество сегментации на предложения в терминах F-меры составило 95,6%, а токенизации — 99,6%.

Особое внимание в НКРЯ уделяется алгоритмам, направленным на наиболее востребованные у пользователей виды разметки: лемматизацию и морфологическую разметку. Качество MyStem не позволяет достаточно точно снимать грамматическую омонимию, из-за чего затруднено использование интегральных видов выдачи, например, группировки примеров по лемме. Переход от MyStem к модели Рубик [Lyashevskaya et al 2023] позволил преодолеть это препятствие. По состоянию на ноябрь 2024 г. моделью Рубик размечены тексты Основного, Газетных и ряда других корпусов. Точность лемматизации составляет 98,79%, определения части речи — 99,03%, определения полного набора морфологических признаков — 97,27%. Помимо этого, применение модели Рубик позволило разметить синтаксическую структуру предложений, что сделало возможным использование условий на синтаксические связи между словами, особенно важных при поиске коллокаций. Точность определения связей без учёта их типа составляет 95,08%, а с учётом — 93,64%.

В Основном корпусе НКРЯ возможен поиск по морфемной структуре слов. При этом во внутреннем словаре, опирающемся на словарь Словарь морфем русского языка [Кузнецова, Ефремова 1986] содержится около 85 тысяч разборов, тогда как в Основном корпусе встречается более 300 тысяч уникальных лемм. Так как ручное пополнение словаря представляется нецелесообразным, было принято

решение пополнить его автоматически. В ходе экспериментов мы определили, что лучших результатов удаётся добиться при помощи ансамбля свёрточных нейронных сетей [Sorokin, Kravtsova 2018]. Дополнив модель рядом лингвистически мотивированных правил, мы достигли доли полностью верных разборов, равной 93,6%, а доли точно определённых морфем — 98,1%.

Помимо внутритекстовой разметки, в НКРЯ присутствует и большое количество метатекстовой разметки. К сожалению, эта разметка в значительной степени неполна. Мы разработали ряд моделей для автоматической доразметки тематики и типа текстов для корпуса региональных СМИ, а также разметки тональности и жанра для корпуса «Социальные сети». Все эти модели представляют собой дообученную на специальных выборках модель RuRoBERTa [Zmitrovich et al 2023]. Появление дополнительной разметки позволяет точнее задавать требующийся пользователю подкорпус.

Таким образом, обширное использование инструментов на базе машинного обучения позволило предоставить пользователям НКРЯ богатые возможности по формированию подкорпусов и использованию более точных и сложных поисковых запросов. Ведётся работа по разработке новых инструментов: разметке ключевых слов, снятию семантической омонимии, генеративной лемматизации на базе больших языковых моделей, генерации толкований и других. Мы рассчитываем расширить существующие сценарии использования НКРЯ и создать новые.

Литература

Кузнецова А. И., Ефремова Т. Ф. Словарь морфем русского языка // Москва : Русский язык, 1986.

Савчук С. О., Архангельский Т. А., Бонч-Осмоловская А. А., Донина О. В., Кузнецова Ю. Н., Ляшевская О. Н., Орехов Б. В., Подрядчикова М. В. Национальный корпус русского языка 2.0: новые возможности и перспективы развития // Вопросы языкознания. 2024. №2. С. 7–34.

Lyashevskaya O., Afanasev I., Rebrikov S., Shishkina Y., Suleymanova E., Trofimov I., Vlasova N. Disambiguation in context in the Russian National Corpus: 20 years later. Компьютерная лингвистика и

интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог». 2023. №2. С. 307–318.

Qi P. et al. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020.

Segalovich I. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine // International Conference on Machine Learning; Models, Technologies and Applications. 2003.

Sorokin A., Kravtsova A. Deep Convolutional Networks for Supervised Morpheme Segmentation of Russian Language // Artificial Intelligence and Natural Language. Springer International Publishing. 2018. С. 3–10.

Zmitrovich D., Abramov A., Kalmykov A., Tikhonova M., Taktasheva E., Astafurov D., Baushenko M., Snegirev A., Shavrina T., Markov S., Mikhailov V., & Fenogenova A. A Family of Pretrained Transformer Language Models for Russian // ArXiv, abs/2309.10931. 2023.

Ю.В. Николаева
(Москва, Россия)
МГУ им. М.В. Ломоносова, Национальный исследовательский университет «Высшая школа экономики»
julianikk@gmail.com

ВОЗМОЖНЫЕ НАПРАВЛЕНИЯ РАБОТЫ В ОРГАНИЗАЦИИ И АННОТАЦИИ МУЛЬТИМОДАЛЬНЫХ КОРПУСНЫХ ДАННЫХ

В докладе предлагается обсудить существующее положение дел с мультимодальными корпусами, достижения и трудности, а также возможные направления дальнейшей работы. Представляется, что доступ к централизованному хранению видеоданных и их аннотаций был бы оптимальным способом сохранить появляющиеся записи и в то же время дать доступ к ним для других исследователей. Это же могло бы способствовать решению существующих трудностей касательно разметки таких данных. Представляется, что несовершенство ручных методов разметки и отсутствие унифицированных подходов не должно служить принципиальным препятствием к этому.

Ключевые слова: жестикуляция, просодия, мультимодальный корпус, МУРКО, RUPEX

В коммуникации лицом к лицу есть две основные модальности (в психологическом значении этого слова): аудиальная для адресата (голосовая для говорящего) и визуальная для адресата (кинетическая для говорящего и жестикулирующего). В первом случае выделяют два канала — вербальный (собственно речь) и просодический (неречевые звуки; среди самых значимых признаков здесь можно назвать высоту основного тона и ее изменение, в том числе его диапазон и скорость; темп; амплитуду или громкость; паузацию; тембр — особенности фонации, такие как придыхание или скрипучий голос). К визуальным/кинетическим каналам относят в первую очередь жесты рук, сопровождающие или замещающие речь, а также жесты головы; помимо этого, сюда можно включить движения плеч, корпуса, ног, мимические выражения, направление взгляда, расстояние между участниками коммуникации и их взаимное положение. Отдельно следует отметить данные движения глаз, полученные с помощью айтрекеров и также представляющие ценный материал для исследователей коммуникации.

При том что реальная коммуникация подразумевает одновременное использование всех этих каналов, создание такого корпуса, который включал бы их полное описание — чрезвычайно трудоемкая задача, поскольку выполняется вручную (автоматическое распознавание движений до сих пор оставляет желать лучшего; автоматическое распознавание просодии и вывод просодической транскрипции также далеки от реализации.

При всем этом мы видим много попыток создать мультимодальные корпуса, размеченные с приемлемой степенью последовательности (Red Hen Lab как самый известный пример). Проблема в том, что почти все они остаются закрытыми, быстро становятся недоступными или про них знает очень малый круг исследователей. В этом плане для русского языка сделано очень многое — МУРКО в составе НКРЯ и RUPEX открыты, достаточно объемны, с подробной разметкой и продолжают развиваться. Менее удачный пример для русского языка — RAMAS, который был доступен некоторое время по запросу. Надо признать, что в несколько лучшем положении находятся корпуса и словари для жестовых языков.

Как представляется, причин нынешних трудностей с мультимодальными корпусами может быть несколько.

Трудоемкость сбора и описания данных сильно ограничивает потенциальный объем корпуса. Вместе с тем надо признать, что даже небольшие подборки видео были бы интересны исследователям. Развитие нейросетей позволяет предположить, что уже в ближайшие годы появятся инструменты для автоматического распознавания и описания мультимодальных данных, и наличие этих данных, размеченных хоть сколько-нибудь последовательно, несомненно, ускорило бы этот процесс.

Отсутствие представления о том, что считать оптимальной разметкой (даже в случае жестикуляции) подводит к вопросу о том, как могло бы выглядеть представление мультимодального корпуса. Как пример — сплошная аннотация записи в RUPEX или выдача кликстов в МУРКО. Вероятно, любые шаги в этом направлении способствовали бы успешной дискуссии о принципах такой разметки.

С другой стороны, частая проблема состоит в том, что существенная часть собранного материала оказывается неразмеченной. Предположу, что даже предварительные и неполные разметки могли бы быть полезны другим исследователям. Другой важный аргумент —

огромный труд, затраченный разметчиками, можно использовать при обучении нейросетей.

Еще одно препятствие к созданию открытых мультимодальных корпусов — отсутствие технических возможностей у их создателей для хранения большого объема данных. Представляется, что существование «хаба», куда можно было бы выгрузить свои видео и разметки, было бы решением для этой проблемы.

Отдельно могут возникнуть трудности этического порядка — насколько допустима публикация видеозаписей в открытом доступе. Однако существует практика сбора согласия (и его отзыва), и в наше время люди гораздо меньше беспокоятся о том, что их лицо попало на видеокамеры, чем 20 лет назад, а запись пересказа фильма, игры с ребенком или попыток говорить на родном языке — не то, что вызывает беспокойство у людей, которые опасаются утечки их личных данных.

Говоря об открытых мультимодальных корпусах, надо признать, что русскому языку повезло. Создание НКРЯ оказалось большим прорывом и в этом направлении тоже: в него входит мультимедийный корпус (МУРКО). В целом работа НКРЯ налажена настолько, что исследования для других языков можно проводить с помощью параллельных корпусов в НКРЯ. Возможно, он мог бы стать таким хабом, который смог бы приютить мини-корпуса, создаваемые другими исследователями. Очевидным образом, это поднимает вопрос о технических и организационных трудностях, в качестве примера можно упомянуть сбор согласий участников записи или техническую проверку подгружаемых видео. Очевидным образом, современные представления о «правильном» корпусе расходятся с тем, что может получиться в результате такой зонтичной организации: частичные, несовместимые и несравнимые друг с другом разметки, разные подходы к ним и очень разные явления, которые оказались в фокусе. Чтобы сделать этот шаг в сторону доступа большего числа исследователей к мультимодальным данным, вероятно, нам придется поменять наши представления о правильном и достойным в таких разметках и смириться с неполными описаниями, ошибками и неразрешимыми противоречиями в подходах. Именно это могло бы позволить нам прийти к полному описанию накопленных записей (с учетом возможностей нейросетей), совершенствованию и унификации описаний и в целом большему пониманию того, какие явления важны в реальном общении и как они между собой взаимосвязаны.

С.И. Переверзева
(Москва, Россия)
Российский государственный гуманитарный университет
P_Sveta@hotmail.com

ГЛУБОКАЯ РАЗМЕТКА ЖЕСТОВ В МУЛЬТИМЕДИЙНОМ РУССКОМ КОРПУСЕ (МУРКО): ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ (2016–2024)

Доклад представляет собой краткий обзор практических и методологических проблем, с которыми сталкиваются в своей работе исследователи, в течение последних 8 лет выполняющие глубокую жестовую разметку Мультимедийного русского корпуса (МУРКО). Для некоторых из этих проблем предлагаются возможные пути решения. Намечаются перспективы дальнейшей работы над жестовой аннотацией МУРКО.

Ключевые слова: мультимедийный корпус, жест, разметка, семантика, художественные фильмы.

Мультимедийный русский корпус (МУРКО) был создан Е.А. Гришиной на базе Национального корпуса русского языка около 14 лет назад. Аннотирование входящих в его состав клипов выполняется автоматически, полуавтоматически и вручную [Кудинов, Гришина 2009]; примером ручного аннотирования служит разметка жестов. До 2016 г. разметку жестов выполняла Е.А. Гришина; ею было размечено 6 художественных фильмов общим объёмом ≈34,5 тыс. слов (≈0,6% от объёма МУРКО). С 2016 г. по инициативе Е.В. Рахилиной черновой разметкой новых клипов занимается С.И. Переверзева совместно со студентами ВШЭ и РГГУ (за прошедшие 8 лет над разметкой работало около 30 чел.) под общим руководством С.О. Савчук. К настоящему моменту почти закончена черновая разметка первой серии фильма «Тот самый Мюнхгаузен» (1979 г., реж. М. Захаров) объёмом ≈5000 слов и ≈2000 жестов.

Задача жестового разметчика состоит в том, чтобы для каждого клипа установить соответствие между наблюдаемым жестом и его возможным значением. Основным ориентиром, в силу требования единообразия, служит уже имеющаяся разметка: «Если при разметке какого-то одного явления принято неточное, если не сказать неправильное решение, то именно это решение, а никакое другое,

должно быть принято и для всех остальных аналогичных явлений» [Гришина 2009: 208–209]. При этом принципы, которым следовала Е.А. Гришина при жестовом аннотировании, нигде не описаны в явном виде, что ставит перед нынешними разметчиками не только проблему выбора между несколькими вариантами аннотации, но и теоретическую задачу реконструкции научного метода. Приведём лишь несколько примеров.

1. В разметке Е.А. Гришиной каждому вхождению жеста в некотором клипе приписывалось только одно значение. Однако жесты по своей природе контекстно многозначны или многофункциональны, и разметчик часто вынужден выбирать: означает жест «поднять брови» удивление или вопрос? «Посмотреть вдаль» — это задумчивость или сосредоточенность? На некоторые из таких вопросов мы попытались дать ответ — см. работу [Переверзева и др. 2019] о выборе между значениями «подтверждение» и «подчеркнуть эмфазу» для иллокутивно независимого кивка и доклад [Вахранёв, Зуева, Переверзева 2020] о выборе значения для жестов, которые могут быть одновременно аналогом речевого действия и выражением эмоции, — но чаще всего подобные решения принимаются субъективно (т. е. так, как кажется правильным большинству из 4-5 человек, работающих над разметкой в данный момент).

2. Принцип единообразия вынуждает разметчиков следовать уже сложившейся традиции. Так, в МУРКО представлены физиологические жесты (жесты, «употребление которых определяется физиологией человека» [Кудинов, Гришина 2009: 252]), например «вздрогнуть» (от холода) или «сморщиться» (от неприятного ощущения). Однако Г.Е. Крейдлин и его соавторы, чьей концепции в основном следовала Е.А. Гришина, считают, что «движениями, а не жестами, будут почесывание (когда чешется), подергивание, вызванное непроизвольным сокращением мышц, движения человека, отгоняющего комаров, гримасы боли» [СЯРЖ 2001: 17]; более того, сама же Е.А. Гришина в своей монографии отмечает, что «приравнять чихание к указательному жесту нам кажется неправильным» [Гришина 2017: 9].

Дальнейшая работа в области жестовой разметки планируется в нескольких направлениях.

1. Поиск и исправление ошибок в ранее созданной разметке МУРКО, а также проверка экспертами черновой разметки новых клипов.

2. Расширение тематики и жанров размечаемых произведений: помимо художественных фильмов планируется аннотирование ток-шоу, дебатов, научно-популярных телепрограмм и докладов на научных конференциях.

3. Создание нового интерфейса выдачи поисковых запросов. Для уже размеченных 6 фильмов были вручную установлены соответствия между жестами персонажей и сопутствующими словами, однако эти соответствия пока недоступны для пользователей корпуса: существующий интерфейс просто не предусматривает такой возможности.

4. Создание рабочего инструмента, позволяющего автоматически сравнивать по частоте употребления жесты и их значения, представленные в МУРКО. Раздел, отражающий статистику жестов и их значений и позволяющий объективно разграничить наиболее и наименее употребительные жесты, мог бы быть полезен не только для исследователей, но и для разметчиков: вопрос о том, сколько новых жестов и значений допустимо вводить при разметке определённого количества клипов, до сих пор не решён.

5. Создание удобной программы для коллективной работы разметчиков. Программа, разработанная М.С. Кудиновым (см. [Кудинов, Гришина 2009]), предназначена для одного разметчика и технически не очень проста в использовании.

6. Создание инструкции для разметчиков, позволяющей сделать выбор при разметке многозначных и синонимичных жестов.

Литература

Вахранёв А. Ю., Зуева А. В., Переверзева С. И. Глубокая разметка Мультимедийного русского корпуса. 2020 год. Доклад на научной конференции «Слово и жест», посвященной памяти Е.А. Гришиной («Гришинские чтения»), 8.02.2020, ИРЯ им. В.В. Виноградова, г. Москва.

Гришина Е. А. Мультимедийный русский корпус (МУРКО): проблемы аннотации // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 175–214.

Гришина Е. А. Русская жестикуляция с лингвистической точки зрения (корпусные исследования). М.: ЯСК, 2017. 744 с.

Кудинов М. С., *Гришина Е. А.* Инструменты полуавтоматической разметки для мультимедийного русского корпуса (МУРКО) // 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. 248–261.

Переверзева С. И., *Ермолаева Н. А.*, *Зуева А. В.*, *Слепак Е. А.* De profundis: проблемы глубокой разметки МУРКО и пути решения // Труды института русского языка им. В.В. Виноградова. 2019. № 21. С. 319–325.

СЯРЖ 2001 — *Григорьева С. А.*, *Григорьев Н. В.*, *Крейдлин Г. Е.* Словарь языка русских жестов. М.; Вена: Языки русской культуры; Венский славистический альманах, 2001. 256 с.

З.Ю. Петрова, Н.А. Фатеева
(Москва, Россия)
Институт русского языка им. В.В. Виноградова РАН
zoyap@mail.ru, nafata@rambler.ru

ИСПОЛЬЗОВАНИЕ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА ПРИ СЛОВАРНОМ ОПИСАНИИ СИСТЕМЫ МЕТАФОР И СРАВНЕНИЙ РУССКОЙ ЛИТЕРАТУРЫ В ЕЕ ДИНАМИКЕ

В работе рассматривается методика пополнения и уточнения базы данных «Материалов к словарю метафор и сравнений русской литературы XIX-XXI вв.» с помощью обращения к Национальному корпусу русского языка, который позволяет выявить состав и динамику образных параллелей на большем массиве данных и прибегнуть к дедуктивному методу описания системы компаративных тропов, предполагающему априорное заполнение «пустых клеток» в семантических полях образов сравнения.

Ключевые слова: метафора, сравнение, словарь, системный подход, эволюция, НКРЯ.

Исследование тропеического уровня языка русской художественной литературы, проводимое в рамках «Материалов к словарю метафор и сравнений <…>» [2000-2021] уже около тридцати лет, подтверждает продуктивность системного подхода к семантическим преобразованиям, в основе которых лежит отношение подобия. Если определить каждый компаративный троп как пару <x, y>, где x — предмет сравнения (tenor, по определению Ричардса), а y — образ сравнения (vehicle, по Ричардсу), то можно сделать вывод, что и x, и y являются элементами семантических полей, которые образуют систему, описываемую в терминах идеографической классификации лексики. Эта система образных параллелей находится в постоянном развитии, как и язык художественной литературы в целом, и задача исследователя состоит в описании динамических процессов в системе компаративных тропов.

Первоначально системная организация метафор и сравнений и эволюция этой системы выявлялась индуктивным способом, из анализа собранного материала. В работе Н.А. Кожевниковой [1995] опреде-

лены основные направления эволюции образных средств с учетом семантических отношений между ними, среди которых отмечаются родо-видовое отношение, отношение «X — совокупность, множество X», «целое — часть» и некоторые другие. Словарь поэтических образов Н.В. Павлович [2007] суммирует накопленные автором данные о крупных классах тропеических конструкций в русской литературе XVIII-XX вв.

На основе исследований Н.А. Кожевниковой и ее обширной картотеки в конце 1990-х годов была начата работа над лексикографическим проектом «Материалы к словарю метафор и сравнений русской литературы XIX-XX вв.», в основу которого был положен указанный выше принцип семантического поля. При группировке материала по выпускам было решено взять за основу семантические поля образов сравнения. Собранный из художественных текстов материал был разделен на поля «Птицы», «Звери», «Насекомые», «Рыбы и другие обитатели моря», «Пресмыкающиеся», «Растения», «Камни», «Металлы», «Ткани, изделия из тканей» и ряд других, менее объемных полей. Для группировки предметов сравнения использовался единый семантический классификатор, созданный по принципу идеографического членения лексического состава языка (авторы учитывали имеющиеся на тот момент классификации); при этом важно, что этот классификатор создавался на основе конкретного имеющегося в картотеке тропеического материала.

Второй принцип группировки языкового материала — по времени написания или первой публикации произведения, из которого взят каждый тропеический контекст, позволил показать развитие образных параллелей во времени.

С 2000 по 2017 гг. были изданы 5 томов словаря, созданных по указанным выше принципам. При группировке тропов по образам сравнения в каждом выпуске индуктивным способом строилось их семантическое поле исходя из основных семантических отношений. Сплошная выборка тропеических контекстов из стихотворных и прозаических произведений около 500 авторов и упорядоченность материала по времени дала возможность получить как приблизительные частотные характеристики тропов, так и показать развитие образных параллелей во времени, появление в них новых элементов,

связанных с традиционными элементами указанными выше семантическими отношениями.

Появление Национального корпуса русского языка дало возможность, во-первых, выявить состав и динамику образных параллелей на большем массиве данных, соответственно, с большей достоверностью, и, во-вторых, прибегнуть к дедуктивному методу, который предполагает априорное заполнение «пустых клеток» в семантических полях образов сравнения. Рабочая методика пополнения базы данных словаря компаративных тропов следующая: полученные при составлении словаря семантические классы образов сравнения тропов пополняются недостающими лексическими единицами соответствующих классов в идеографических словарях (наиболее удобен для этого Семантический словарь под ред. Н.Ю. Шведовой), привлекается и материал из терминосистем соответствующих научных областей, если в идеографическом словаре отсутствуют какие-то единицы. По каждой из единиц класса проводится поиск как по Основному корпусу (подкорпус Художественные тексты), так и по Поэтическому корпусу. Из полученных множеств контекстов отбираются употребления лексической единицы в метафорических и сравнительных конструкциях, конструкциях параллелизма. Этими употреблениями пополняются списки контекстов употребления уже имеющихся в словаре образов сравнения и в некоторых случаях характеризуемых ими предметов сравнения; классы образов сравнения пополняются новыми элементами. Таким образом уточняются сведения о составе фрагментов системы метафор и сравнений, об их изменении во времени.

В докладе иллюстрируется методика работы с Национальным корпусом русского языка для внесения уточнений и дополнений в базу данных «Материалов к словарю метафор и сравнений XIX-XXI вв.». на примере одного небольшого фрагмента семантической категории образов сравнения «Звери», а именно «Пресмыкающиеся». Дополненную и уточненную базу данных можно будет в дальнейшем использовать для пополнения базы данных электронного Интерактивного словаря метафор и сравнений, который был создан нами в 2020 г. [Козеренко, Петрова, Ребецкая, Фатеева — электронный ресурс].

Литература

Кожевникова Н.А. Эволюция тропов // Очерки истории языка русской поэзии XX в. Образные средства поэтического языка и их трансформация. М.: Наука, 1995. С. 6–79.

Козеренко А.Д., Петрова З.Ю., Ребецкая Н.А., Фатеева Н.А. Интерактивный словарь компаративных тропов русской литературы XIX-XX вв. [Электронный ресурс]. URL: https://ruslang.ru/interaktivnyy-slovar-komparativnykh-tropov-russkoy-khudozhestvennoy-literatury-xix-xxi-vv

Материалы к словарю метафор и сравнений русской литературы XIX–XXI в. Вып. 1–6. / Сост. Кожевникова Н.А., Петрова З.Ю. (вып. 1–5); Петрова З.Ю., Фатеева Н.А. (вып. 6). М.: Издательский дом ЯСК, 2000–2021.

Павлович Н.В. Словарь поэтических образов: на материале русской художественной литературы XVIII-XX вв. Т. 1. М.: УРСС, 2007. 848 с.; Т.2. М.: УРСС, 2007. 869 с.

В.И. Подлесская
(Москва, Россия)
Институт языкознания РАН
podlesskaya@iling-ran.ru

МУРКО И ЕГО БРАТЬЯ: ЗВУКОВЫЕ КОРПУСА В ИССЛЕДОВАНИЯХ УСТНОЙ РЕЧИ

Доклад посвящен роли МУРКО и ряда малых аудио- и видеокорпусов в исследовании устной речи. Показано, что корпуса с просодической разметкой могут предоставить важный материал для решения не только дескриптивных, но и общетипологических задач. Демонстрируется, что в русском устном дискурсе обнаруживаются явления, которые не вписываются в привычный типологический профиль русского языка, в том числе, конструкции с дислоцированным топиком, редупликация, сериализация, цепочечные структуры.

1. Задачи. Скачок, который произошел в исследованиях устной речи благодаря электронным корпусам, оказался поистине революционным даже на фоне общих тектонических сдвигов в эмпирической лингвистике в связи с развитием корпусных методов. Пионером и чемпионом среди корпусных ресурсов устной речи безусловно является МУРКО — мультимедийный подкорпус НКРЯ. Десятилетиями основным источником данных для исследователей устной речи были так называемые «блокнотные» записи. МУРКО позволил перевести документирование устной речи в новый формат — цифровые аудио- и видеозаписи, доступные для многократного воспроизведения и обработки с помощью компьютерных анализаторов речи. Внушительный объем и изощренные возможности поиска делают МУРКО незаменимым инструментом. Вместе с тем, при таком объеме данных невозможно отразить в транскрипте извлеченного фрагмента дискурса все многообразие его характеристик — особенно просодических. Поэтому постепенно возникают небольшие по объему электронные корпуса устной речи, главный козырь которых — подробная просодическая разметка. Я имею счастье принадлежать команде, силами которой разрабатываются такого рода корпуса с просодической разметкой. В докладе ставятся две задачи: (1) очень кратко охарактеризовать эту группу корпусов; и (2) проиллюстрировать возможности корпуса с просодической разметкой на примере его использования для решения одной из множества задач, а именно показать, что в русском устном дискурсе обнаруживаются

явления, которые не вписываются в привычный типологический профиль русского языка.

2. Малые братья МУРКО. Речь пойдет о трех ресурсах, объединенных единой программой исследования устного дискурса и единым форматом используемой просодической транскрипции. Обзор этих ресурсов представлен в работе [Коротаев и др. 2023]; принципы и формат просодической транскрипции описаны в [Кибрик, Подлесская 2009, Kibrik et al. 2020 a, b]. Первый ресурс — электронная коллекция «Рассказы о сновидениях и другие корпуса звучащей речи» (http://spokencorpora.ru), он содержит инициированные монологи (рассказы по картинкам, рассказы на заданную тему), представленные в формате аудиозаписи и синхронизированного с ней транскрипта. Второй ресурс — «Рассказы и разговоры о грушах», новаторский проект по анализу мультиканального дискурса (см. [Кибрик, Федорова 2018]; https://multidiscourse.ru). Этот корпус включает серию записей естественной коммуникации между несколькими участниками, с синхронизированной вокальной и кинетической аннотацией этих записей, а также регистрацией движений глаз. Третий ресурс — пилотный проект корпуса «Что я видел» [Коротаев и др. 2024], включающий устные и письменные версий нарративов двух повторяющихся сюжетов: рассказа о запомнившемся сновидении и яркой истории из жизни. Испытуемыми выступили взрослые (15 — 59 лет) носители русского языка. У части из них было диагностировано паническое расстройство; остальные составили контрольную группу. Всего 155 текстов от 39 испытуемых. Данные корпуса «Что я видел» используются в качестве иллюстраций в данном докладе.

3. Русский устный как «экзотический» язык. Привлечение данных просодии, дает возможность не только ставить новые задачи, но и по-новому подойти к задачам, имеющим солидную историю. Одна из них — составление типологического портрета того или иного языка. Данные просодически размеченного корпуса, в частности, позволили показать, что в устном русском языке обнаруживаются, по крайней мере, следующие феномены, которые не принято включать в стандартные типологические характеристики русского языка. Среди них: два семантических типа редупликации с двумя разными просодическими паттернами — длительное итерируемое событие (два акцентных пика) и незначительное непродолжительное итерируемое событие (один акцентный пик); левая дислокация топикальной группы с сохранением ее прономинальной копии в основной клаузе; сериальные конструкции с интегрированным интонационным

контуром; так называемые цепочечные структуры — длинные цепочки предикаций, из которых только последняя маркирована, как завершенная. В докладе будут приведены иллюстрации этих явлений, в том числе, соответствующие транскрипты и визуализации движений тона в программе PRAAT. Мы постараемся показать, что привлечение данных устной речи и особенно, просодии может существенно пополнить наши представления о языковом многообразии.

Литература

Кибрик А.А., Подлесская В.И. (ред.). Рассказы о сновидениях: корпусное исследование устного русского дискурса. М.: Языки славянских культур. 2009.

Кибрик А.А., Федорова О.В. An empirical study of multichannel communication: Russian Pear Chats and Stories. Психология. Журнал Высшей Школы экономики. 2018. 15 (2). С.191–200.

Коротаев Н.А., Литвиненко А.О., Подлесская В.И. «Вам и не снилось!»: прошлое, настоящее и будущее «рассказов о сновидениях» // Язык как он есть: Сб. ст. к 60-летию Андрея Александровича Кибрика / Ред.-сост. Т.И. Давидюк, И.И. Исаев, Ю.В. Мазурова, С.Г. Татевосов, О.В. Федорова. — М.: Буки Веди. 2023. С. 432−437

Коротаев Н. А., Д. А. Панышева, Е. А. Неверова, В.И. Подлесская. Корпус «Что я видел» как инструмент анализа панического дискурса // «Слово и жест». Научная конференция, посвященная памяти Е. А. Гришиной («Гришинские чтения»). Москва, 8 февраля 2024 г. Материалы конференции. / Отв. ред. С. О. Савчук. − М.: Институт русского языка им. В. В. Виноградова РАН. 2024. С. 32−37

Kibrik A.A., Korotaev N.A., Podlesskaya V.I. Russian spoken discourse: Local structure and prosody // S. Izre'el, H. Mello, A. Panunzi, and T. Raso (eds.). In search of basic units of spoken language: A corpus-driven approach. Netherlands: John Benjamins Publishing Company. 2020a. P. 35–76.

Kibrik A.A., Korotaev N.A., Podlesskaya V.I. 2020b. The Moscow approach to local discourse structure: An application to English // S. Izre'el, H. Mello, A. Panunzi, and T. Raso (eds.). In search of basic units of spoken language: A corpus-driven approach. Netherlands: John Benjamins Publishing Company. 2020b. P. 368–382.

Поляков А.Е.
(Москва, Россия)
pollex@mail.ru

ЛЕММАТИЗАТОРЫ ДЛЯ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА

1. Определения.

Лемматизатор — это программа, выполняющая лексико-грамматический анализ текста, который включает следующие задачи:

1) Токенизация — разбиение текста на элементарные знаки (токены) и определение их типа (слово, знак препинания, число, тег разметки).

2) Лемматизация — приведение словоформы к лемме (словарной форме) и определение ее грамматических признаков (число, падеж, время, наклонение...), причем для одного слова возможно несколько вариантов разбора, включая ноль.

3) Построение гипотез для нераспознанных (несловарных) слов на основе существующих (словарных).

4) Снятие омонимии в случае нескольких вариантов (если возможно).

Грамматическая модель — формальное описание словоизменения языка, включающее два компонента: грамматический словарь и таблица парадигм.

Грамматический словарь — список лексем с грамматической информацией, включая:

1) основа с указанием чередований;

2) постоянные признаки лексемы (часть речи, род, одушевленность...);

3) код словоизменительного типа (парадигмы).

Словоизменительный тип (парадигма) — набор флексий, общий для некоторого множества лексем, который задает соответствие между грамматическими значениями и соответствующими им формами.

2. Постановка проблемы.

НКРЯ включает тексты, относящиеся к разным периодам и формам русского языка: древнерусский (11-15 в.); старорусский (15-17); церковнославянский (17-18); современный русский (19-21 в.);

диалектный и т.д. Для разных периодов применяются разные программы и методы морфологического анализа. Для современного русского работает автоматическая лемматизация (mystem), но для получения корпуса со снятой омонимией требуется ручная правка. Для церковнославянского применяется простой лемматизатор на основе словаря словоформ. Для древнерусского возможна только ручная лемматизация.

Помимо чисто языковых различий, тексты одного периода могут иметь графико-орфографические особенности. Так, тексты 17-19 века могут быть набраны как в старой орфографии (с ятем и ером), так и в модернизированной. Церковнославянские тексты часто имеют различную орфографию в зависимости от времени и места издания.

3. Лемматизатор для современного русского языка.

Программа mystem (Яндекс) создана на базе словаря Зализняка, который был приведен в формальный вид, пригодный для морфологического анализа. Она включает грамматический словарь с указанием чередований и список парадигм, выдает все варианты для словарных слов и умеет строить гипотезы для незнакомых (несловарных) слов. Но есть проблемы:

1) Словарь и парадигмы зашиты в программу и их невозможно изменить или дополнить.

2) Программа выдает маловероятные разборы для многих частотных слов: *для* = деепр. от *длить*, *при* = импер. от *переть*, *ли* = мера длины, *он* = имя буквы, *их* = междометие. Пришлось писать специальные фильтры, которые отсекают лишние варианты.

3) Продуктивные словообразовательные модели учтены в словаре только частично. Там есть некоторые субстантивированные прилагательные и местоимения (*все, это, каждый, больной, старое, белые, черные*...), но реально почти любое прилагательное может субстантивироваться. Аналогично, отадъективные наречия могут образовываться почти от любого прилагательного, но в словаре есть только самые частотные, а остальные получают разбор как краткая форма ср. рода.

4) Гипотезы для незнакомых слов выдаются в непредсказуемом порядке, правильный вариант часто оказывается в конце списка. Для

разных форм одного слова часто выдаются абсолютно разные наборы лемм:

Мономах => пр.мн. от *моном* или *монома* (~~ *домах, дамах, проблемах*)

Мономаха => *мономах* или *мономаха* (~~ *монах, росомаха*)

салями => тв.мн. от *саль* или *саля* (~~ *королями, солями*)

Программа разбивает слово на основу+флексию и выбирает самые частотные (в словаре) комбинации: *сал+ями*, *моном+ах* (*дом+ах* vs. *монах*), но часто неудачно.

После обучения на снятом корпусе программа mystem стала работать значительно лучше:

1) из словаря убрали совсем маловероятные разборы;

2) программа научилась снимать часть омонимии по контексту.

4. Лемматизатор для дореформенной русской орфографии.

Лингвистические процессоры, ориентированные на современный русский язык, непригодны для анализа старых текстов (18–19 века). Они правильно анализируют формы, совпадающие с современными (*рука, новый, милость*), но считают ошибкой старые формы (*домъ, домъ, новаго, милостію, безсильныя, съузить, ходити*).

Mystem частично понимает старую орфографию путем приведения ее к современной (*і=и, ѣ=е, ѳ=ф, ѵ=и*), а также знает некоторые старые флексии (*-аго, -яго, -ыя, -ия*). Эта функция добавлена по нашей просьбе для проекта ФЭБ (http://feb-web.ru), где многие тексты даны в старой орфографии. Для адаптации программы к языку 18–19 века нужна серьезная переделка словаря и грамматических таблиц, однако код программы и словарей закрыт.

Пришлось писать свой лемматизатор с возможностью гибкой настройки и адаптации к различным вариантам русского языка и орфографии. Основные принципы:

1. Вся конкретно-языковая информация (леммы, основы, парадигмы, флексии, граммемы) не фиксирована в коде программы, а вынесена в отдельные таблицы.

2. Программа преобразует текст во внутреннее (нормализованное) представление, которое унифицирует орфографические различия:

1) заменяет старые буквы на современные эквиваленты (*і=>и, ѣ=>е, ѳ=>ф, ѵ=>и*);

2) отсекает конечный *-ъ*;

3) заменяет начальные *без-/в(о)з-/из-/низ-/раз-/роз-/ч(е)рез-* => *-с* перед глухими (NB);

4) заменяет некоторые недопустимые сочетания букв на правильные, в частности:

– убирает лишние еры внутри слова (*съиграть=>сыграть, съузить, съэкономить, съагитировать, съзади, близълежащий*);

– нормализует гласные после шипящих (*чюжой=>чужой, жывой=>живой*).

После нормализации многие "неправильные" формы становятся вполне правильными и получают разборы, хотя часть информации теряется. Так, совпадают квази-омонимы (ѣ–е): *слѣзъ–слезъ, сѣлъ–селъ, свѣдѣніе–сведеніе, Вѣна–вена, морѣ–море*. Также пропускаются явные ошибки: *ѣлка=елка, ѳѵзіка=физика, історія=история*.

В основе программы словарь Зализняка и модель словоизменения современного русского языка, но добавлены старинные или неучтенные формы (флексии):

1) Адъективные флексии (*-аго/-яго, -ыя/-ія*).

2) Усеченные формы прилагательных (*красна/о/ы/у*), которые совпадают с краткими формами, но имеют и другие падежи (*добра молодца, красну девицу*).

3) Особые формы местоимений (*ея, онѣ, однѣ, однѣхъ*).

4) Творительный падеж 3-го склонения на *-ію* (*милостію, помощію*).

5) Сравнительная степень на *-яе/-яй* (*сильняе, скоряе*) и *-ея* (*сильнея*).

6) Вариант частицы *-ся* после гласных (*валюся, валилася*), который употребляется также в современном языке.

7) Деепричастия совершенного вида от основы презенса (*прийдя, увидя, взгромоздясь*), которые вполне употребительны в современном языке.

8) Глагольные флексии *-ти* и *-ши* (*ходиши, ходити*), которые характерны для 18-ого века.

В грамматических таблицах эти формы имеют специальные пометы (old1, old2, old3), и их можно включать/отключать в настройках.

Действующая модель лемматизатора написана на Javascript и работает прямо в браузере.

http://dic.feb-web.ru/russian/parser/parser.htm

Реальный лемматизатор написан на Python и используется для анализа текстов XVII–XVIII веков.

5. Лемматизатор для церковнославянского корпуса.

Церковнославянский (ЦС) в русской среде воспринимается не как иностранный язык, а как архаичный книжный русский, где есть особая лексика и устаревшие формы (аорист, имперфект). Русский литературный язык буквально пропитан ЦС лексикой, даже некоторые грамматические формы имеют ЦС происхождение:

1) активные причастия (*делающий, сделавший* vs *делая, сделав*)

2) превосходная степень (*сильнейший, высочайший* vs *сильнее, выше*)

ЦС корпус включает богослужебные тексты, в основном изданные в XIX–XX веках, и несколько более архаичных (XVIII век). Язык этих текстов достаточно унифицирован и может быть проанализирован автоматически. Но есть проблемы:

1) Сложная графико-орфографическая система, включая дополнительные буквы и сложные правила их употребления (є/e, о/o/w, и/i/v, ф/ѳ, з/s, ѯ, ѱ, титла, буквотитла, ударения, придыхание).

2) Сложная система словоизмения, особенно для глаголов и прилагательных.

3) Орфографическая и грамматическая нестабильность, когда одна и та же форма может писаться разными способами, иногда даже в соседних строках.

ЦС лемматизатор был задуман как развитие лемматизатора для старой орфографии. Были разработаны:

1) Правила перевода текста во внутреннее (унифицированное) представление.

2) Таблицы парадигм на базе описаний в грамматиках ЦС языка.

3) Мини-словарь с парадигмами для обучения модели.

4) Предсказатель для анализа неизвестных словоформ на основе существующих.

Мы обработали программой список словоформ и проверили полученные результаты. Оказалось, что предсказатель выдает много лишних разборов, причем правильные разборы часто оказываются в конце списка. ЦС словоизменение значительно сложнее русского и некоторые формы дают негативный эффект для работы предсказателя. Например, аорист и имперфект типа *дѣлахъ, дѣла* вторгаются в склонение существительных и порождают там массу лишних вариантов. Титла и буквотитла искажают графическую форму слова и не позволяют привести его к лемме (*агг҃лъ=>ангелъ, апс̃лъ=>апостолъ, гл҃ати=>глаголати*). Поэтому значительную часть словоформ пришлось исправлять вручную, заодно проверяя их в реальном корпусе, чтобы выбрать правильный вариант.

Грамматики ЦС языка описывают идеальную картину, которая часто не соответствует реальному состоянию текстов. Существует масса орфографических и словоизменительных вариантов, в том числе явные русизмы (*лукѣ, маркѣ, ученикамъ, учениками* вм. *-wмъ, -и*).

В результате нам удалось получить список словоформ с разборами, который настроен на существующий корпус. При расширении корпуса неизбежно появятся новые словоформы и список придется расширять.

Словарь словоформ можно посмотреть здесь:

http://dic.feb-web.ru/slavonic/dicgram/index.htm.

Лемматизатор для ЦС корпуса выделяет слово из текста, преобразует во внутреннее представление, находит слово в списке и подставляет разборы во входной текст.

Результат работы лемматизатора можно посмотреть в корпусе:

http://ruscorpora.ru/search-orthlib.html.

Н. А. Ребецкая
(Москва, Россия)
ИРЯ им. В.В. Виноградова РАН
n.reb@mail.ru

БАЗА ДАННЫХ СЛОВАРЯ ЯЗЫКА ПУШКИНА. КОРПУСНЫЕ ИССЛЕДОВАНИЯ

В работе рассматриваются результаты исследований по материалам базы данных Словаря языка Пушкина, созданной в рамках реализуемого в Институте русского языка им. В.В. Виноградова проекта «Электронизация Словаря языка Пушкина». Инструментарий базы данных позволяет получить быстрый доступ ко всем словарным статьям и их компонентам, в том числе тем, которые в печатном издании представлены лишь шифрами. В поисковую систему включены опции поиска по различным пометам, с помощью которых составители уточняют тип значения и окраску слов. Жанрово-хронологический принцип разбиения таблиц базы данных позволяет выделить разные по жанрам и периодам подкорпусы для сравнительного анализа помеченных словоупотреблений.

*Ключевые слова***:** Словарь языка Пушкина; база данных; корпусные исследования; переносные словоупотребления; фразеологизмы.

Основной задачей проекта «Электронизация Словаря языка Пушкина», реализуемого в Институте русского языка им. В.В. Виноградова, является создание на основе лемматизированного конкорданса полноценных словарных статей для всех словоформ (за исключением самых частых слов), входящих в Полное собрание сочинений А.С. Пушкина в 16 томах (СС). Для этих целей конкорданс был преобразован в базу данных (БД), поля таблиц которой соответствуют разделам словарной статьи в печатном издании. Инструментарий БД позволяет осуществлять поиск по лексеме, отбирать фразеологические сочетания, а также контексты, где слова употреблены в переносном значении или имеют стилистические пометы.

В настоящее время в Базе данных присутствует лексический материал из 8 первых томов СС, в состав которых входит лирика трех периодов (1813-1817 гг., 1817–1825 гг., 1826–1836 гг.), цикл стихотворений «Песни западных славян», сказки, поэмы двух периодов

(1817–1824 гг. и 1825–1833 гг.), роман в стихах «Евгений Онегин», драма и проза.

Жанрово-хронологический принцип структурирования таблиц БД дает возможность выделить разные по жанрам и периодам подкорпусы для сравнительного анализа того или иного типа информации.

На основе имеющегося материала с помощью инструментария БД был осуществлен ряд корпусных исследований элементов словарных статей, в частности помет, с помощью которых составители уточняют тип значения и окраску слов. Рассмотрим эти исследования.

Авторы Словаря языка Пушкина разметили все случаи употребления слов в переносных значениях и во фразеологических сочетаниях. Данные пометы стали объектом нашего первого исследования. Для сравнительного анализа по жанрам было выделено 3 подкорпуса: лирика, драма и проза — и подсчитана доля переносных словоупотреблений в каждом (Рис. 1).

Рис. 1. Распределение долей словоупотреблений с пометой «переносное» по подкорпусам лирики, драмы и прозы.

Доля словоупотреблений с переносным значением в лирике (5,2% от общего числа с/у) превышает долю таковых в драме (1,8%) почти в 3 раза, а в прозе (0,6%) более чем в 8 раз. Этот количественный показатель отражает характерную черту поэзии — использование металогического приема, с помощью которого автор в стихотворном тексте выражает свои мысли, ощущения, желания. Метафоры служат

для создания эмоционального эффекта, художественных образов и представлений об окружающем мире. Особенно высока доля слов с пометой «переносное» в стихах второго тома (1817–1825), сочиненных в наиболее плодотворный в творческом плане период южной ссылки — период расцвета романтизма.

В пределах тех же подкорпусов было проведено исследование отмеченных составителями фразеологических сочетаний. На рис. 2 показано распределение долей словоупотреблений, входящих в сочетания, по подкорпусам лирики, драмы и прозы.

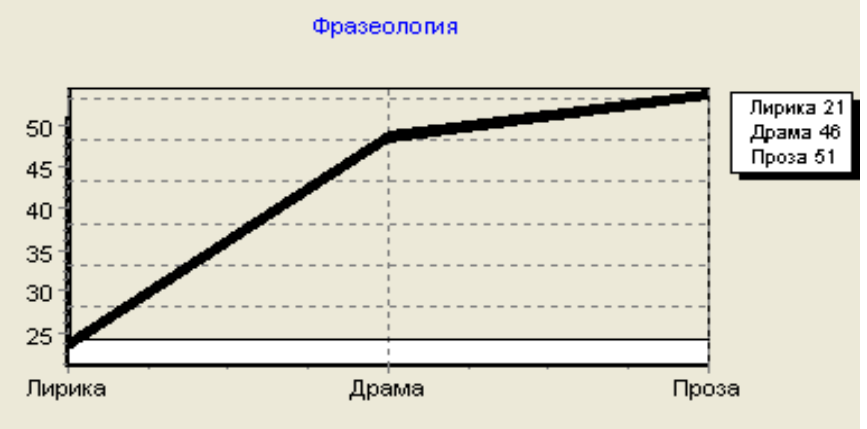


Рис. 2. Распределение долей словоупотреблений, входящих во фразеологические сочетания, по подкорпусам лирики, драмы и прозы

Наблюдаем здесь обратную по сравнению с предыдущим графиком (рис.1) картину: доля фразеологизмов резко увеличивается в драматических произведениях и достигает максимума в прозе (в лирике 2,1%, в драме 4,6%, в прозе 5,1%).

Анализ графиков на рисунках 1 и 2 позволяет сделать следующий вывод: фразеологические сочетания являются характерным средством художественной выразительности в прозе Пушкина, тогда как переносные употребления можно считать маркером художественных средств в поэзии. Драма с ее смешанным форматом — в поэтический текст включены прозаические фрагменты с высокой долей разговорных оборотов — занимает промежуточное положение, но при этом по составу художественных приемов она ближе к прозе.

Те же словарные элементы были исследованы на материале выборок из подкорпусов поэм двух периодов — раннего (1817–1824 гг.) и позднего (1825–1833 гг.) (рис. 3).

Рис. 3 Распределение долей словоупотреблений с пометами «переносное» и «сочетание» по подкорпусам поэм раннего и позднего периодов

График отражает увеличение доли фразеологических сочетаний (темная линия) и уменьшение доли переносных словоупотреблений (светлая линия) к концу творчества. В произведениях раннего периода, среди которых «Руслан и Людмила», «Кавказский пленник», «Бахчисарайский фонтан», где образы героев окружены атмосферой отвлеченных романтических символов и метафор, больше переносных значений, доля народно-поэтического стиля, основного источника фразеологизмов, невелика.

В поэмах позднего периода («Граф Нулин», «Полтава», «Домик в Коломне», «Медный всадник») пушкинский стиль всё теснее сближается с сокровищами родного слова, всё больше проникается духом народного русского языка. Пушкин использует здесь формулы простонародного слога и оборотов просторечия.

Данное исследование подтвердило основное направление в поиске средств художественной выразительности у Пушкина — стремление поэта «проникнуть в формы народной поэзии, ввести простонародный, песенный стих в литературу», сочетая подобные художественные приемы с другими литературными стилями.

Третье исследование касается стилистических помет в СЯП.

Надо заметить, что наиболее распространенные стилистические пометы, такие как *книжное, высокое, разговорное, просторечное*, за исключением некоторых отметок типа «в речи крестьянина», «в речи лиц из простого народа», в словаре отсутствуют. Тем не менее, составители включили некоторые стилистические пометы в словарные статьи для выделения оттенков значения, и большинство этих помет указывают эмоциональную окраску слова: *шутливое, ироническое, каламбурное, бранное, уничижительное*. Встречается также помета, указывающая на стилистическую ограниченность употребления слова в литературном языке — *народно-поэтическое*. Для анализа был выбран подкорпус Онегин, наиболее полно представленный в иллюстративном материале словарных статей СЯП.

Рис. 4. Распределение словоупотреблений со стилистическими пометами в подкорпусе Онегин

Представлены следующие пометы: *шутливое, ироническое, каламбурное, народно.-поэтическое, уничижительное.* Как видно из графика, количество шутливых и иронических помет в «Евгении Онегине» преобладает. Это подтверждает выводы, сделанные исследователями стиля Пушкина: роман, несмотря на глубину и серьёзность, весь пронизан юмором и иронией, включая бытовые описания, авторские отступления, саму фигуру главного героя, его

воспитание, снобизм… Здесь Пушкин верен своему принципу лёгкой подачи сложного содержания.

Таким образом, на основе корпусных исследований информации в словарных статьях Словаря языка Пушкина с применением статистического анализа были выявлены характерные художественные приемы различных литературных жанров и раскрыта картина эволюции образных средств, использованных Пушкиным на протяжении всего его творчества.

Эта работа стала возможной благодаря созданной на основе оцифрованного шестнадцатитомного собрания сочинений базе данных, включающей словарные статьи с полной информацией для каждого словоупотребления, в том числе и тех, что в Словаре языка Пушкина обозначены лишь шифром, а также благодаря информационно-поисковой системе, которая позволяет извлекать необходимую информацию из подкорпусов, относящихся к различным типам художественных произведений и временным периодам жизни поэта. Добавим, что в опции поисковой системы БД помимо рассмотренных в статье входят следующие элементы: *ласкательное; мифология; официальное; перифрастическое; ремарка; церковное; цитата; эвфемистическое; название; в значении существительного; заглавие.* Выборки словоупотреблений с этими пометами откроют обширные горизонты для исследования различных граней творчества Пушкина.

Литература

Ребецкая Н. А. О некоторых аспектах электронизации Словаря языка Пушкина. Журнал Смоленского центра квантитативной филологии, 2023, № 1 (5). С. 64-71.

Ребецкая Н. А., Шайкевич А. Я. Электронизация «Словаря языка Пушкина» // Труды Института русского языка им. В. В. Виноградова, 2024, №1. С. 49-57.

Л. В. Рычкова
(Гродно, Республика Беларусь)
Гродненский государственный университет имени Янки Купалы
rychkova@grsu.by

НКРЯ КАК ИСТОЧНИК ДЛЯ ФОРМИРОВАНИЯ ЭКСПЕРИМЕНТАЛЬНО-ДОКАЗАТЕЛЬНОЙ БАЗЫ НАУЧНЫХ ИССЛЕДОВАНИЙ МАГИСТРАНТОВ-ЛИНГВИСТОВ

На примере ряда диссертаций, выполненных магистрантами, обучавшимися по специальности «Теоретическая и прикладная лингвистика», показан потенциал НКРЯ как источника для формирования экспериментально-доказательной базы лингвистических научных исследований. Помимо основного корпуса, были использованы возможности таких модулей НКРЯ, как параллельный корпус, корпус региональной и зарубежной прессы, исторический и панхронический модули.

Ключевые слова: корпусные технологии, модули НКРЯ, экспериментально-доказательная база лингвистических исследований, корпусная грамотность.

Корпусная лингвистика как учебная дисциплина исчезла из учебных планов специальностей первой ступени высшего образования Республики Беларусь вместе со специализацией «Компьютерная лингвистика», а затем и направлением «Компьютерное обеспечение». Лишь в 2019 году дисциплина «Корпусная лингвистика» была рекомендована для включения в учебные планы подготовки магистрантов по специальности «Теоретическая и прикладная лингвистика».

В Гродненском государственном университете имени Янки Купалы впервые на постсоветском пространстве начал преподаваться курс корпусной лингвистики, задолго до того, как в систему высшего образования страны была введена магистратура. Преподавание осуществлялось кафедрой общего и славянского языкознания, существовавшей на филологическом факультете вплоть до его реорганизации в 2017 году, поэтому, даже при отсутствии соответствующей дисциплины в учебных планах, корпусные технологии использовались в учебном процессе при обучении другим лингвистическим дисциплинам. Потому закономерным продолжением многолетней практики использования корпусных технологий в

университете стало обучение магистрантов рекомендованной для них дисциплине «Корпусная лингвистика». Отметим, что подготовка магистрантов по специальности «Теоретическая и прикладная лингвистика», которым преподается данная дисциплина, осуществляется не на филологическом факультете, а на факультете истории, коммуникации и туризма кафедрой перевода и межкультурной коммуникации.

Прежде всего, стоит отметить, что портал НКРЯ представляет собой уникальную базу для обучения корпусной лингвистике и широко используется нами при обучении видам корпусной разметки, типологии корпусов, видам корпусных запросов и дополнительных опций получения корпусных данных, а также форматам выдач и особенностям работы с ними. Важно, что меню портала позволяет легко получить доступ к справочным материалам, публикациям, посвященным НКРЯ, а также к другим корпусам. Обучение магистрантов основам корпусной лингвистики формирует у них достаточный уровень корпусной грамотности, что позволяет им использовать возможности НКРЯ не только для формирования экспериментально-доказательной базы диссертационных исследований, но и для проведения самих исследований. Приведем примеры тем некоторых из выполненных с использованием возможностей различных модулей НКРЯ магистерских диссертаций: «Взаимодействие белорусского и русского языков в текстах газет Гродненщины (на основе корпусных данных)», «Прилагательные с семантикой цвета в произведениях русской литературы для детей: дидактический аспект», «Лингвокультурологический аспект передачи русскоязычных соматизмов средствами китайского языка (на материале параллельного корпуса НКРЯ)», «Система значений грамматических категорий рода и падежа русского существительного в преломлении на итальянский язык (на материале параллельного корпуса НКРЯ)», «Аномальные и искаженные формы в коммуникации на русскоязычных форумах (на материале данных основного модуля Национального корпуса русского языка)», «Англоязычные вкрапления в текстах русскоязычных СМИ: прагматический аспект (на материале корпуса региональной и зарубежной прессы НКРЯ)», «Специфика перевода русскоязычных nomina feminina на английский и француз-

ский языки (на материале параллельных корпусов)», «Глобализация в аспекте лексической тональности (на материале корпусных данных)».

Рассмотрим подробнее два из вышеназванных исследований, которые, на наш взгляд, отличает оригинальность использования возможностей НКРЯ. Так, при выполнении диссертационного исследования на тему «Прилагательные с семантикой цвета в произведениях русской литературы для детей: дидактический аспект» магистранткой с использованием возможностей семантической разметки основного модуля НКРЯ был осуществлен поиск цветообозначений в специально сформированном подкорпусе детской литературы и осуществлена классификация колоративов с точки зрения их дидактического потенциала для различных возрастных категорий дошкольников. Выявление обучающего потенциала детской литературы для освоения цветообозначений дошкольниками позволило магистрантке разработать комплекс развивающих заданий с использованием цветообозначений, функционирующих в целевом подкорпусе. Ею также была разработана методика применения разработанного комплекса заданий и проведена его пилотная апробация, показавшая высокую эффективность в освоении дошкольниками колоративов.

При выполнении диссертационного исследования на тему «Специфика перевода русскоязычных nomina feminina на английский и французский языки (на материале параллельных корпусов)» магистранткой были использованы возможности нескольких модулей НКРЯ. На первом этапе исследования с целью выявления специфичных для русской лингвокультуры nomina feminina она обратилась к историческому и панхроническому модулям НКРЯ, что позволило составить референтный список целевых для ее исследования номинаций, который был на следующем этапе уточнен с использованием данных о частотности конкретных nomina feminina в основном корпусе НКРЯ. На следующем этапе исследования также с использованием инструментария основного корпуса был проведен анализ частотности использования целевых номинаций в текстах различных авторов. Затем был проведен анализ состава русско-английского и русско-французского параллельных подкорпусов для выявления автора и произведения, содержащего наибольшее количество контекстов, в которых встречаются специфичные для

русской лингвокультуры nomina feminina, и его переводов на английский и французский языки. Таким произведением стал роман Ф. М. Достоевского «Преступление и наказание», который отличается своеобразным языком с точки зрения использования уменьшительно-ласкательных слов, стилистически-маркированной лексики, авторских неологизмов, а также слов, сознательно используемых автором с нарушением языковых норм. Такой подход позволил магистрантке обосновать оптимальность выбора текста именно данного художественного произведения в качестве источника для формирования экспериментально доказательной базы исследования. При проведении самого исследования магистранткой дополнительно были использованы возможности семантической разметки параллельного корпуса, в частности, для выявления тональности (положительной и отрицательной оценки) nomina feminina в тексте оригинала и специфики ее отражения в переводах, а также аугментативных и диминутивных существительных, относящихся к исследуемой категории.

Таким образом, возможности, предоставляемые НКРЯ, позволяют осуществлять проведение многоплановых актуальных лингвистических исследований, результаты которых находят непосредственное применение на практике. Все вышеназванные диссертационные исследования проводились магистрантами по заявкам учреждений и организаций г. Гродно.

Сабольч Янурик
(Будапешт, Венгрия)
Будапештский университет им. Лоранда Этвеша
janurik@yahoo.com

ВОПРОСЫ ПРИМЕНЕНИЯ КОРПУСНЫХ ДАННЫХ В ЛЕКСИКОГРАФИЧЕСКОМ ОПИСАНИИ АНГЛИЙСКИХ ЗАИМСТВОВАНИЙ В РУССКОМ ЯЗЫКЕ

В докладе обсуждаются перспективы использования корпусных данных при описании английских заимствований в русской лексикографии. Выдвигается тезис о том, что сопоставление информации, обнаруживаемой в словарных и корпусных материалах, позволяет уточнить различные аспекты процесса заимствования этих иноязычных лексем, в том числе время их вхождения в русский язык, а также особенности их графической и морфологической адаптации. На основе анализа некоторых контекстов, в которых встречаются англицизмы в НКРЯ, рассматриваются возможности включения результатов корпусных исследований в лексикографическое представление этих слов.

Ключевые слова: корпусные данные, английские заимствования в русском языке, первая фиксация слова, графическая адаптация, морфологическая адаптация, частотность, вариативность

И. А. Смаль, Д. А. Морозов
(Новосибирск, Россия)
НП «НКРЯ»
vanasmal@mail.ru, morozovdm@gmail.com

РАЗМЕТКА СЕМАНТИКИ В СИСТЕМЕ НКРЯ

В работе представлен обзор текущего состояния семантической разметки НКРЯ. Описаны существующие проблемы, включающие в себя недостаточно структурированную систему семантических помет, а также отсутствие механизмов снятия омонимии. Также предложены первые шаги на пути к их решению, включающие в себя составление новой системы помет, а также реализацию алгоритмов машинного обучения для автоматического снятия омонимии.

Ключевые слова: семантика, снятие омонимии, машинное обучение, НКРЯ, обработка естественного языка

Задача определения семантики является одной из классических задач обработки естественного языка и имеет множество разновидностей. В системе НКРЯ присутствует как оценка семантики текста целиком, которая выражена разметкой тематики и ключевых слов [Савчук и др. 2024], так и разметка семантики отдельных слов [Рахилина и др. 2009]. В этом докладе рассматривается только вторая разновидность задачи.

Решать задачу определения семантики отдельных слов можно различными способами. Первый способ заключается в описании семантики каждого отдельного слова, например, как в толковом словаре. А второй — в составлении общей структуры помет и сопоставлении каждому слову некоторого набора этих помет. Эти наборы в зависимости от алгоритма могут быть из заранее определенного множества наборов, а могут быть произвольными.

В НКРЯ сейчас применяется второй вариант, в котором возможные наборы помет для каждой леммы определены заранее, причем семантическая омонимия не снимается: один набор (наиболее вероятный) обозначается как «основной», а остальные — как «дополнительные». Из-за этого существенно снижается ценность поиска по семантическим пометам. Кроме того, если слово отсутствует в словаре (например, слово «видеокарта»), то оно не

получит никаких помет, а у присутствующих слов иногда неполный список значений (например, у слова «Наполеон» не существует набора помет, соответствующих значению названия торта).
Помимо этого, сама система помет в НКРЯ обладает рядом недостатков. Во-первых, в ней присутствует много редких помет: более 1 тыс. из них представлены в словаре одним-двумя словами. Вторым существенным недостатком является непостоянная система иерархии, в которой слову может присуждаться более специфичная помета, но не присуждаться более общая (например, слову «плохой» приписывается помета «ev:negative», но не приписывается помета «ev»). В-третьих, семантика перемешана с информацией о происхождении слова: например, слово «дынный» имеет помету dt:food, которая приписывается словам, образованным от слов с пометой t:food, но не имеет самой пометы t:food. Все это делает поиск по семантическим пометам сложным для неподготовленных пользователей.

Мы предлагаем подход, который состоит из двух частей: во-первых, обновлённая упрощённая система семантических помет (улучшающая как однородность природы разметки, так и баланс классов), во-вторых, алгоритм снятия семантической омонимии и автоматического расширения разметки на слова, отсутствующие в словаре.

Новая система основана на атомарных пометах, и иерархия в ней осуществляется с помощью добавления новых, уточняющих помет, а не использованием напрямую более узких помет. Так, например, помета «ev:negative», которая раньше заключала в себе как наличие некоторой оценки, так и уточнение, что эта оценка негативная, будет соответствовать наличию двух отдельных помет «ev» и «negative», а новая помета «time» является объединением того, что раньше было 26-ю различными пометами, так или иначе связанными со временем. Благодаря этому поиск по общим классам будет осуществляться проще и не будет требовать полного перебора всех более узких классов при составлении запроса.

Для снятия семантической омонимии мы предлагаем алгоритм, представляющий собой расширенную версию ранее представленного алгоритма Рубик [Lyashevskaya et al 2023]. Исходный алгоритм представляет собой композицию из кодировщика и

трёх декодировщиков, где результат одного из декодировщиков (размечающего морфологические признаки) может быть переиспользован для двух других (при лемматизации и разметке синтаксических связей). Показано, что исходный алгоритм позволяет достаточно точно решать задачу снятия грамматической омонимии, вследствие чего целесообразным представляется исследование возможности модификации алгоритма для работы и с семантической омонимией. Наш подход заключается в разработке четвёртого декодировщика, использующего как скрытое векторное представление, так и результаты работы остальных декодировщиков. Разработанный декодировщик решает задачу многометочной (multilabel) классификации.

Литература

Рахилина Е. В., Кустова Г. И., Ляшевская О. Н., Резникова Т. И., Шеманова О. Ю. Задачи и принципы семантической разметки лексики в НКРЯ // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы / отв. ред. В.А. Плунгян. СПб.: Нестор-История, 2009. С.215–239.

Савчук С. О., Архангельский Т. А., Бонч-Осмоловская А. А., Донина О. В., Кузнецова Ю. Н., Ляшевская О. Н., Орехов Б. В., Подрядчикова М. В. Национальный корпус русского языка 2.0: новые возможности и перспективы развития // Вопросы языкознания. 2024. №2. С. 7–34.

Lyashevskaya O., Afanasev I., Rebrikov S, Shishkina Y., Suleymanova E., Trofimov I., Vlasova N. Disambiguation in context in the Russian National Corpus: 20 years later. Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог». 2023. №2. С. 307–318.

Т. П. Соколова
(Москва, Россия)
Московский государственный юридический университет имени О. Е. Кутафина
tsokolova58@mail.ru

ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА ДЛЯ ПРОИЗВОДСТВА СУДЕБНОЙ ЛИНГВИСТИЧЕСКОЙ ЭКСПЕРТИЗЫ

В докладе на конкретных примерах рассматриваются возможности использования ресурсов обновленного и модернизированного Национального корпуса русского языка для производства судебной лингвистической экспертизы, в том числе для выведения контекстуального значения слов, еще не зафиксированных словарями современного русского языка, для определения лексической сочетаемости таких слов, а также показаны возможности использования корпусов НКРЯ для производства лингвистической экспертизы спорных наименований.

Ключевые слова: НКРЯ, корпусная лингвистика, судебная лингвистическая экспертиза, нейминговая экспертиза

Созданный 20 лет назад и непрерывно развивающийся Национальный корпус русского языка (НКРЯ) стал неотъемлемым инструментом лингвистических исследований в целом [Рюкова 2024] и лингвистической экспертизы, как особой прикладной деятельности в правовом поле, в частности. Корпусный анализ с конкретными примерами решения задач судебной лингвистической экспертизы (определения функциональных особенностей слова, словосочетания, предложения, определения семантических особенностей указанных единиц и их стилистических особенностей) включен в раздел семантической методологии методического пособия по судебной лингвистической экспертизе [Судебная лингвистическая экспертиза 2023: 54–67]. Однако в результате фундаментальной реконструкции и модернизации платформы НКРЯ в 2020-2023 гг. возможности обновленного и существенно обогащенного речевыми произведениями разных стилей и жанров корпуса, снабженного усовершенствованным инструментарием поиска, репрезентации, индексации текстов, их особой разметкой [Савчук и др. 2024: 7–34], стали значительно шире указанных в пособии.

Именно НКРЯ служит надежным источником выявления семантики новых слов и выражений, еще не зафиксированных словарями современного русского языка. Так, требования, предъявляемые к заключению эксперта-лингвиста, не позволяют ссылаться на многочисленные неофициальные онлайн-словари субстандартной лексики при определении лексического значения слов наркотического дискурса, который постоянно пополняется новыми единицами. Эксперту-речеведу необходимо обосновать «выводное значение» новой лексемы, и сделать это позволяет работа с НКРЯ. Например, в спорном тексте, поступившем на экспертизу, выявлено слово *кладмен,* которое отсутствует в нормативных словарях современного русского языка, однако зафиксировано в онлайн-словаре субстандартной лексики AntiSlang.ru. Однако это общедоступный интернет-словарь, в который «сами пользователи имеют возможность вносить слова самостоятельно», следовательно, источник ненадежный и нелегитимный. Поиск по лемме «кладмен» в Газетном корпусе и в подкорпусе «Социальные сети» НКРЯ, дал следующие результаты: найдено 16 примеров употребления словоформ леммы «кладмен», в том числе с поясняющим контекстом: «*человек, делающий закладку наркотиков»*; *«сотни кладменов — пехотинцев наркобизнеса, делающих «закладки» товара, — разносят сотни и тысячи доз наркотиков»* и др. Таким образом, с помощью лексико-семантического и контекстуального анализа эксперт вывел лексическое значение слова *кладмен* — ‘человек, распространяющий наркотические вещества путем «закладок» / «потайных кладов»’, ‘участник сети наркоторговли, синоним «закладчик»’. Принадлежность лексемы *кладмен* к семантическому полю «наркотики» подтверждается также «Облаком слов» — ближайшими семантическими ассоциатами слова, которые выводятся автоматически НКРЯ (эта новая опция позволяет включить в заключение эксперта наглядную иллюстрацию вхождения нового слова в указанное семантическое поле).

НКРЯ представляет широкие возможности выявления лексической сочетаемости в контексте речевых произведений разных стилей и жанров, что весьма востребовано в неймингoвой экспертизе — особом роде судебной лингвистической экспертизы, объектом которой становятся коммерческие наименования [Соколова 2019: 196-207]. Например, чтобы ответить на поставленный перед экспертом

вопрос «Каково значение слова «Останкино» в современном русском языке, связано ли значение слова «Останкино» с телевидением и радиовещанием вообще, с названием района в г. Москва?», необходимо использование не только данных лексикографических источников (ономастических словарей), но и корпусов НКРЯ для уточнения референтного значения наименования «Останкино» в диахроническом аспекте. Данные корпусов НКРЯ позволили выявить и подтвердить статистическими данными изменение коллокаций слова *Останкино*: от исторических (*деревня, музей-усадьба, парк, район*) к современным (*телевидение, телевидение и радиовещание, телекомпания, телевизионный центр*). Сопоставление данных основного и газетного корпусов Национального корпуса русского языка показывает изменение частотности референтов слова *Останкино*, а именно увеличение числа словоупотреблений со значением ‘телевидение и радиовещание’, уменьшение числа словоупотреблений со значением ‘район’, ‘музей-усадьба’, ‘парк’. Если ранее (до строительства телецентра и телебашни) ориентиром и доминантой была усадьба *Останкино* с парком и музеем, то с 1960-х гг. телецентр (и телебашня) *Останкино* становятся доминантой не только района *Останкино*, но и города, а вследствие метонимического переноса слово ОСТАНКИНО получает значение — ‘телевидение и радиовещание’. На основе анализа данных корпусной базы НКРЯ эксперт сделал вывод о широком употреблении слова ОСТАНКИНО, его общеизвестности в силу регулярной встречаемости в текстах литературного русского языка и современных СМИ. Преобладающим референтом слова ОСТАНКИНО является *телевидение и радиовещание*, которое осуществляет *Телевизионный центр «Останкино», телерадиокомпания «Останкино».*

Таким образом НКРЯ является не только легитимным источником необходимой эксперту-речеведу языковой информации, но и важным инструментом корпусного анализа в судебной лингвистической экспертизе.

Литература

Национальный корпус русского языка. [Электронный ресурс]. URL: https://ruscorpora.ru/

Рюкова А. Р. Корпусно-ориентированные исследования языка: краткий обзор достижений и трудностей // Russian Linguistic Bulletin.

2024. № 1 (49). [Электронный ресурс]. URL: https://rulb.org/media /articles/10132.pdf

Савчук С. О., Архангельский Т. А., Бонч-Осмоловская А. А., Донина О. В., Кузнецова Ю. Н., Ляшевская О. Н., Орехов Б. В., Подрядчикова М. В. Национальный корпус русского языка 2.0: новые возможности и перспективы развития // Вопросы языкознания, 2024, № 2. С. 7–34.

Соколова Т. П. Нейминговая экспертиза как особый род судебной лингвистической экспертизы. Медиалингвистика, 2019. № 6(2). С. 196–207.

Судебная лингвистическая экспертиза. Теория, методики, практика: методическое пособие / Под ред. В.О. Кузнецова, А.М. Плотниковой. — М.: ФБУ РФСЦЭ при Минюсте России, 2023. 232 с.

AntiSlang.ru [Электронный ресурс]. URL: https://antislang.ru/about

В. Д. Соловьев
(Казань, Россия)
Казанский федеральный университет
maki.solovyev@mail.ru

ПОЛНОТА И СБАЛАНСИРОВАННОСТЬ: СОПОСТАВИТЕЛЬНЫЙ АНАЛИЗ НКРЯ И GOOGLE BOOKS NGRAM[1]

В статье обсуждается проблема сбалансированности диахронических корпусов. Описан метод выявления несбалансированности в тех корпусах, состав которых не известен. Метод продемонстрирован на примере подкорпусов корпуса Google Books Ngram. Приведено сопоставление Основного корпуса НКРЯ и Google Books Ngram с точки зрения репрезентативности. Даны рекомендации по учету возможной несбалансированности при проведении лингвистических и социологических исследований на базе таких корпусов.

Ключевые слова: диахронические корпусы, сбалансированность, репрезентативность, Google Books Ngram

Проблема сбалансированности корпусов обсуждается более 30 лет и представляется одной из наиболее сложных. НКРЯ позиционируется как “наиболее сбалансированный (в нем представлены тексты самых разных жанров приблизительно в той пропорции, в которой с ними сталкивается обычный носитель языка)” (https://ruscorpora.ru/page/faq/). Однако представляется трудным объективно определить, с чем носитель языка сталкивается реально (особенно 100, 200, 500 лет назад). При создании Google Books Ngram (далее, GBN, https://books.google.com/ngrams/), реализован подход сплошного представления в корпусе всех опубликованных текстов, хранящихся в библиотеках (которые удается найти и сканировать с хорошим качеством). Кажется, что это хороший метод обеспечения сбалансированности, в предположении, что издательская политика и библиотечная системы сами сбалансированы.

[1] Исследование выполнено за счет гранта Российского научного фонда № 24-18-00570.

Ранее в GBN было выявлено одно существенное нарушение, когда в подкорпус художественной литературы было включено много научных книг. Google оперативно это исправил. Нами обнаружено еще одно важное нарушение сбалансированности, отразившееся на корпусе, начиная с 2000 г., − резкое необоснованное увеличение числа художественных произведений.

Если взять случайным образом слова, характерные для художественной литературы и редко встречающиеся в других жанрах (например, *помада, шорох*), то при высокой стабильности частот этих слов в XX веке мы видим резкий немотивированный всплеск частоты в 2000 г., и затем увеличение их частоты в несколько раз к 2022 г. с всплесками в района 2011 и 2020 гг. (рис. 1). Выполнена проверка для 100 слов по словарю художественной литературы [Частотный 2009], которая привела к таким же результатам. По данным Министерства цифрового развития, связи и массовых коммуникаций никакого увеличения выпуска художественной литературы в этот время не происходит, она держится на уровне 15%. Никакого увеличения частоты этих слов в НКРЯ также не зафиксировано. Единственное возможное объяснение такого синхронного изменения частот слов состоит во включении в корпус GBN большого количества художественной литературы этого периода. По меньшей мере в отношении доли художественной литературы, НКРЯ является много более сбалансированным, чем GBN. Это видно по представленной в НКРЯ диаграмме (https://ruscorpora.ru/corpus/main/stats/chrono) изменения доли художественных книг.

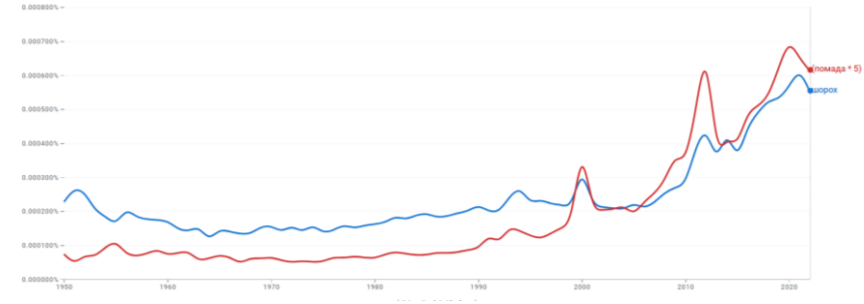


Рис. 1. График частотности слов *помада* (масштабировано) и *шорох* в корпусе GNB без сглаживания

Следует отметить, что аналогична картина наблюдается и для других языков в GNB — увеличение доли художественной литературы в XXI веке. Мотивы подобного решения Google неизвестны. Представляется, что обнаруженный дисбаланс слишком велик и его нужно учитывать в статистических исследованиях, в том числе выполняя отдельно расчеты частотности до 2000 г. и после него.

В аспекте репрезентативности картина иная. За счет объема (русский подкорпус GBN в 200 раз больше основного корпуса НКРЯ), для редких слов в НКРЯ статистика недостаточна для обоснованных умозаключений. Более того, некоторые весьма значимые события жизни общества не находят адекватного отражения в НКРЯ, и в то же время четко представлены в GBN. Например, это касается решения об “ускорении” развития страны, принятого на пленуме ЦК КПСС в 1985 г. График GBN отражает рост частотности этого слова именно с 1985 г. и последующее резкое падение с 1987 г. (рис. 2), когда его сменило слово *перестройка*. Основной корпус НКРЯ никак не фиксирует это важное событие в жизни общества, хотя газетный корпус НКРЯ показывает схожую динамику частотности слова *ускорение*.

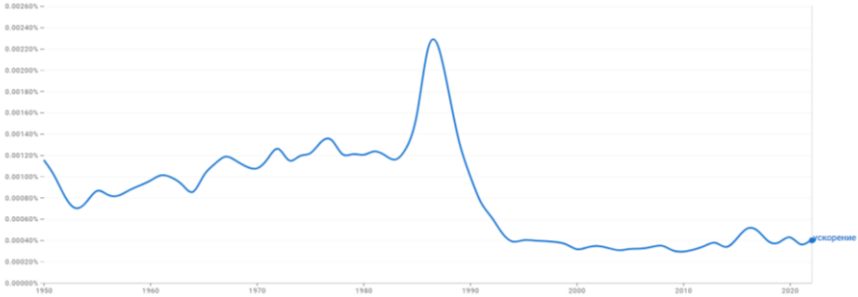


Рис. 2. График частотности слова *ускорение* по GBN

По какой-то причине литература, отражающее это явление, не попала в основной корпус НКРЯ, и будущим исследователям следует иметь в виду возможность такого рода отклонений.

Литература

1. Частотный словарь современного русского языка [Электронный ресурс]. URL: http://dict.ruslang.ru/freq.php.

И. И. Столяров, О. А. Митрофанова
(Санкт-Петербург, Россия)
ООО «Центр искусственного интеллекта МТС», Санкт-Петербургский государственный университет
i.stolyarov@mts.ai, o.mitrofanova@spbu.ru

АВТОМАТИЧЕСКОЕ РАЗГРАНИЧЕНИЕ ОМОГРАФОВ (НА МАТЕРИАЛЕ НКРЯ)

В докладе рассматривается проблема автоматического снятия омографии — актуальная теоретическая и прикладная задача обработки текста. На материале корпуса омографов, составленного на основе НКРЯ, анализируются различные методы для разграничения омографов разных типов. В качестве оптимального решения предлагается два возможных подхода — на основе извлечения лингвистической информации и с использованием моделей семейства BERT для русского языка.

Ключевые слова: лексическая неоднозначность, морфологическая неоднозначность, омография, снятие неоднозначности, снятие омографии, предобработка текста

Доклад посвящён проблеме разграничения омографов — слов одной или разных частей речи, совпадающих в написании, но различающихся произношением и имеющих разные значения [Емельянова 2003]. В русском языке омография может быть связана с разной постановкой ударения (*хло́пок — хлопо́к, учи́теля — учителя́*) и практикой неразличения на письме букв «е» и «ё» (*все* может означать и ‘все’, и ‘всё’). В некоторых случаях возможна комбинация указанных факторов: *бе́рег — берёг, сестры́ — сёстры*. Снятие омографии, т. е. определение правильного варианта произнесения омографа в данном контексте, может рассматриваться как частный случай снятия лексической или морфологической неоднозначности, но вместе с тем эта задача имеет и чисто прикладное значение: так, при синтезе речи по тексту правильная расстановка ударений напрямую влияет на качество синтезированной речи [Соломенник и др. 2013].

Для настоящего исследования был разработан корпус омографов на основе Национального корпуса русского языка. Корпус включает в себя 84 400 контекстов для 56 омографических пар разных типов и одной омографической «триады» (*се́ла — сёла — села́*)? каждый

контекст представлен распространённым предложением с целевым омографом. Корпус выложен в открытый доступ по адресу https://github.com/iv-stoliar/RussianHomographDataset. В докладе описывается процесс отбора и предобработки контекстов, который позволил нам сделать несколько наблюдений об НКРЯ.

На материале собранного корпуса проводится обзор и тестирование существующих решений для снятия омографии: применение автоматического расстановщика ударений RusStress, обращение к тезаурусу RuWordNet для разграничения лексических омографов (*за́мок — замо́к*), использование теггеров и морфоанализаторов (Natasha, RNNMorph, spaCy) для разметки грамматических (*ру́ки — руки́*) и лексико-грамматических (*ве́сти — вести́*) омографов. Полученные в этих экспериментах результаты свидетельствуют о том, что разграничение омографов разных типов по-прежнему представляет сложность для всех инструментов автоматической обработки текста.

В рамках настоящего исследования предлагается два подхода, которые позволяют достичь сопоставимых или более высоких результатов снятия омографии. Первый подход предполагает использование разных видов лингвистической информации, извлекаемой из контекста. Для лексических омографов (*за́мок — замо́к*) алгоритм дизамбигуации основан на определении семантической близости между ближайшим контекстом омографа и рядами слов, отражающих различные семантические связи с каждым из вариантов омографа (гиперонимы, гипонимы и т. д.). В случае грамматических (*ру́ки — руки́*) и лексико-грамматических (*ве́сти — вести́*) омографов предлагается применить методы машинного обучения с использованием морфологических, синтаксических и/или семантических признаков контекста. Также в этих экспериментах был автоматически рассчитан вклад разных признаков в классификацию.

В рамках второго подхода предлагается использовать языковые модели семейства BERT (англ. Bidirectional Encoder Representations from Transformers), которые порождают контекстуализированные эмбеддинги слов и могут успешно применяться для снятия разных видов неоднозначности. В настоящем исследовании мы отталкивались от эксперимента по разграничению омографов в английском языке [Nicolis, Klimkov 2021]. Было протестировано 6 моделей BERT для

русского языка, различающихся архитектурой и обученных на разных данных. Сравнив результаты экспериментов, мы смогли сделать вывод о том, что оптимальным решением (и по скорости работы, и по эффективности) является использование моделей меньшего размера, дообученных на большем объёме материала.

Представляется, что рассмотренные подходы могут быть успешно использованы для снятия других видов неоднозначности, выходящих за рамки омографии, — лексической неоднозначности (для разграничения омонимов *лук, ключ* и т. д.) и морфологической неоднозначности (для разграничения омоформ *стали* от сущ. *сталь* и *стали* от глаг. *стать* и т. д.). Также представленные решения могут найти применение на практике, в частности в системах синтеза речи по тексту.

Литература

Емельянова О. Н. Омонимия и смежные явления // Стилистический энциклопедический словарь русского языка / Отв. ред. М. Н. Кожина. М., 2003. С. 263–267.

Соломенник А. И., Таланов А. О., Соломенник М. В., Хомицевич О. Г., Чистиков П. Г. Оценка качества синтезированной речи: проблемы и решения // Изв. вузов. Приборостроение. Тематический выпуск "Речевые информационные системы". 2013. №2. С. 38–42.

Nicolis M., Klimkov V. Homograph disambiguation with contextual word embeddings for TTS systems // Proceedings of the 11th ISCA Speech Synthesis Workshop (SSW 11), 2021. P. 222–226.

Е.В. Туркина
(Москва, Россия)
Институт русского языка им. В. В. Виноградова РАН
turkina.ketrin@yandex.ru

«ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В МЕДИЙНЫХ И СЕТЕВЫХ ТЕКСТАХ: ЧЕРТЫ КОРПУСНОГО ПОРТРЕТА»

Результаты корпусного исследования представлений о «сильном», «слабом» и «персональном» искусственном интеллекте, отраженных в актуальном медиадискурсе и социальных сетях, позволяют определить в структуре соответствующего тематического поля группы «друг», «враг», «помощник», которым соответствуют коннотации, характеризующие каждую из ипостасей искусственного интеллекта.

Ключевые слова: Искусственный интеллект, медийный и сетевой дискурс, тематическое поле, тематическая группа, потрет выражения, корпусный анализ.

Искусственный интеллект (ИИ) осмысливается и семантизируется в дискурсивном поле (в частности, в медиадискурсе, где обсуждение нейросетей и ИИ стало главным трендом) одновременно с его технологическим развитием.

Если проследить стадии развития ИИ, можно выделить представления о трех его ипостасях, отразившиеся в текстах современных медиа и социальных сетей. На основании данных НКРЯ на первом этапе исследования [Северская, Туркина 2024] в тематическом поле «Искусственный интеллект» были выделены представленные конкретными текстовыми реализациями семантические классы:

- *Сильный ИИ*;
- *Слабый ИИ;*
- *Персональный ИИ*.

Контент-анализ Газетных корпусов НКРЯ позволил выявить и сравнить языковые воплощения представлений о каждой из этих ипостасей.

Объемлющее понятие «искусственный интеллект» без конкретизации одним из прилагательных *сильный*, *слабый*, *персональный*, передающее специфику ИИ и этапность его развития — от *слабого* — к

сильному и *персональному*, присутствует в 2994 текстах и 5335 примерах употреблений Газетных корпусов НКРЯ.

Конкретизация вида ИИ и обозначение этапа его развития происходит за счет расширения контекста, в котором реализуется объемлющее понятие ИИ. Контент-анализ позволил собрать данные и разграничить ипостаси ИИ, определить отношение (коннотации) носителей и пользователей языка к ИИ на его разных этапах развития.

Сильный ИИ характеризуется как «общий» или «универсальный». Он, по замыслу конструкторов, должен быть «человекоподобным», то есть обладать способностью, начав с нуля, вникнуть в любую область и начать ориентироваться в ней подобно тому, как это способен делать человек [Редозубов 2021: 145].

Как друг Сильный ИИ в Газетном корпусе — это «интеллектуальная система управления», «лидер», что «наравне с человеком», «инструмент», при этом он — «хулиганский», «способный на самостоятельные решения», обучение, «компактный и неприхотливый», «в нем нет ни сердца, ни души, ни сострадания», «полноценный». Используется для «управления» в разных сферах.

В социальных сетях Сильный ИИ расценивается как друг-помощник — «компьютер» и «нейросеть», для которой должны быть прописаны чёткие правила использования.

Как враг Сильный ИИ представлен в Газетном корпусе в формах «цифровизации», «социальной алгоритмизации», «системы с признаками разума» и осознается «инструментом полицейского государства», «оружием, способным принимать решение без человека». Сильный ИИ в роли врага в медийной сфере — это также пиар и маркетинговый инструмент, который «начинает приобретать черты очередного хайпа», и помощник террористов, который стал причиной «роста угроз совершения террористических преступлений». Медиадискурс отражает и представления о враждебных действиях ИИ: он «выходит за рамки технологий», «за всё отвечает, избавляет от ответственности», «конкурирует с себе подобными», «будет обладать признаками разума», «захватит мир». А используют его в этом контексте для «манипулирования людьми», «нанесения им увечий».

В соцсетях Сильный ИИ в определенных контекстах позиционируется как «фундаментальная угроза для Человечества».

Слабый ИИ позволяет решать творческие задачи (и чаще всего только одну конкретную). Он абсолютно несамостоятелен, человек его обучает.

В Газетном корпусе он представлен как друг и помощник, об этом «говорит» не только контекст, но и соответствующие глаголы (ИИ «помогает», «может помочь»), предлоги (что-то делается «благодаря» ИИ), предложно-падежные сочетания («с помощью ИИ зарабатывают»; ИИ «приходит на помощь», его усилия «должны быть направлены на помощь человеку»). В соцсетях мы видим такое же использование глагольных форм: ИИ «может помочь», «объединит всех», ему что-то «поручили придумать», можно «отправиться в путешествие с помощью» ИИ.

В языке массмедиа слабый ИИ — это «технологии», «алгоритмы», «система», «просто счётная машина», «цифровой помощник учителя», «функциональный помощник», «очень большое поле для творчества», «инструмент для творцов».

Персональный ИИ «способен» понимать и чувствовать людей, иметь свою «личность» или полностью нашу (примеры: Siri Apple, Яндекс Алиса и т.д.). Глаголы, маркирующие действия, которые совершает ИИ, в Газетном корпусе НКРЯ преимущественно будущего времени, что подтверждает тезис о том, что Персональный ИИ — пока только идея, разработка, цель. В корпусе на это указывают ключевой глагол «сможет» и его синонимы: «мониторинг эмоционального состояния сотрудника с помощью ИИ *станет* обычным делом», ИИ «*будет* определять эмоции и реальные предпочтения аудитории», ученые «*смогут* доработать (ИИ)», ИИ «*сможет* пройти тест Тьюринга», «человек *сможет* обзавестись цифровым двойником», ИИ «*окажется* слишком говорливым»).

Как друг Персональный ИИ — лидер с человеческими качествами, «вычислительная модель», «технология», «робот», «главный драйвер рынка средств человеко-машинных коммуникаций», «нейросеть», «цифровой двойник — виртуальный аналог организма». Как враг — угроза, компьютер, обманщик, который «не станет президентом».

В соцсетях он «обманывает» и «блефует», и это воспринимается как угроза и большая опасность (ср. сообщения о том, как некто

«покончил с собой после общения с девушкой, созданной нейросетью», «привязался к искусственно-созданной собеседнице»). ИИ — это враждебная «нейросеть», «искусственно-созданная собеседница», «холодный» и «расчётливый».

Функция «помощник» для ИИ, судя по контекстам, является базовой, а гипертрофированная акциональность ИИ как технологии превращает его из «друга» во «врага».

Литература

Северская О.И., Туркина Е.В. Искусственный интеллект в современном медиадискурсе: сильный, слабый, персональный // Terra Linguistica. 2024. Т. 15, № 3. С. 74–80. DOI: 10.18721/JHSS.15307

Редозубов А.Д. Формализация смысла. Часть 1 // Онтология проектирования. 2021. Т. 11, №2(40). С.145. DOI: 10.18287/2223-9537-2021-11-2-144-153

Е.В.Филиппова
(Москва, Россия)
Академия ГПС МЧС России
elenefilippova.klachuk@yandex.ru

ТРАНСФОРМАЦИЯ СМЫСЛОВОГО НАПОЛНЕНИЯ ПОНЯТИЯ «МИЛОСТЬ»

В статье рассматривается изменение значения концепта «милость» в советскую эпоху и новейшее время. Автор показывает, что идеология государства оказывает влияние на семантическое наполнение понятия и сферу использования его в речи.

Ключевые слова: религиозная сфера, понятие, толковый словарь, НКРЯ, лингвокультурология.

В современном лингвокультурном сознании русского человека этические понятия занимали немаловажное место, поскольку духовное развитие общества всегда имело свое вербальное отражение в языке. Этические концепты как необходимая составляющая ментальности любого народа определяли ценности народа и устанавливали критерии для разграничения добра и зла. Актуальность нашего исследования обусловлена тем, что в лингвокультуре современного человека наблюдается перемещение некоторых компонентов этических понятий на периферию, в пассивный запас, и тенденция к употреблению таких понятий в контекстах не связанных с морально-нравственными ценностями, а только для создания иронии и сарказма. Целью нашей статьи является описание изменений, происходящих в наполнении понятия «милость» на основе Национального корпуса русского языка (НКР) и определение исторических и культурологических предпосылок, влияющих на трансформацию значений, составляющих объём понятия.

Слово «милость» восходит к древнерусскому слову «милъ, миловати, милость, милостьнъ, милосердиѣ», к старославянскому «милъ» — ‘возбуждающий милосердие, сострадание’; сербохорватскому — ‘ласкать, любимец, милашка’; к греческому μείλιον ‘приятный дар’ [Фасмер); ἐλεεῖν — ‘жалость, вызываемая охватывающим всю душу состраданием’ [Орлов], к еврейским «хен» — ‘благосклонность, расположение к определенному человеку’ и «хесед» — ‘поступок, вознаграждающий кого-либо за преданность и

верность, за оказание помощи' [Ринекер]. Безусловно, эти значения являются калькой из иностранных языков, которые были заимствованы, скорее всего, с целью достоверного перевода богослужебных текстов на русский язык, где милость понималась как дар Бога человеку.

Проанализировав дефиниции этого слова в толковых словарях советского и постсоветского времени (МАН, С.А. Кузнецов, Т.Ф. Ефремова, Н.Ю. Шведова, С.И. Ожегов), мы установили, что слово имеет примерно одинаковое наполнение, которое включает следующие компоненты: 'доброе человеколюбивое отношение'; 'благодеяние, дар'; 'благосклонность, полное доверие'; 'сострадание'. Исключение составляет Большой универсальный словарь РЯ, где слово это отсутствует. Изучение ассоциатов двух самых близких слов «милосердие» и «милость» в НКРЯ показал, что для слова «милосердие» «милость» является ассоциатом, а для «милости» «милосердие» — нет. Для «милости» ассоциатами выступают «благоволение», «снисхождение», «изволение», «благость», «благодеяние», «щедрота», «благосердие», что свидетельствует об оттенке в значении слова «милость». Этот дополнительный компонент восстанавливается на основе дефиниций и примеров из словарей И.И. Срезневского и В. Даля, где примерами выступают цитаты из Священного Писания или выражения со словом «Бог», а основными компонентами являются 'прощение' и 'любовь', что указывает на религиозный контекст бытования этого слова в дореволюционный период.

Подтверждением может служить цитата из святоотеческого наследия Григория Богослова: «Награждай добрых, презирай злых. Но и последним оказывай одну милость: нимало не оскорбляйся их поступками, чтобы великодушием и их со временем сделать добрыми. Прекрасный дар — милость» [Симфония]. Эти важные компоненты значения были намеренно утрачены в словарях советского периода, в связи с тем, что религиозная сфера была объявлена антигосударственной и всеми способами вытеснялась из сознания советского народа. Известно высказывание Ушакова относительно статей в словарях: «История словарей показывает, что каждый из них является классовым отражением своей эпохи».

После возрождения интереса к религиозной жизни возродилось и понятие «милость», о чем нам свидетельствует НКРЯ на

представленных графиках, но, к сожалению, сферой использования этого слова осталась только религиозная сфера, в частности верующие люди: христиане, мусульмане, иудеи. В светском языке наблюдаются изменения. В современном русском языке новейшего времени слово «милость», по данным НКРЯ, функционирует в контекстах, посвященных судебным разбирательствам, политической жизни страны и чаще всего встречается в составе коллокатов: «*оказать, явить, сотворить милость, ждать, просить милости, рассчитывать на милость*», где обязательно должен быть активный субъект (агенс), обладающий большой силой, или большим влиянием или, наоборот, униженный бедный, претерпевающий лишения человек. На месте первых обычно оказываются государственные власти, президент или судебные органы, а на месте вторых — обвиняемые или рядовые обманутые или обездоленные граждане: *Не стоит ждать милости от Пенсионного фонда, свой капитал на старость разумнее сколачивать самому»; «Среди сгоревших строений всего два было застраховано, остальным собственникам приходится рассчитывать лишь на милость государства, которое по закону вовсе не обязано выдавать компенсацию собственникам», — констатировал эксперт».* Милость здесь можно получить только от субъектов, обладающих силой и могуществом, а не добродетелями, милующими по жалости, а не по любви.

Искоренение веры из сознания русского человека повлекло и постепенное перемещение слова «милость» в пассивный запас приобретения им пометы «устар.». Но в языке закрепилось несколько устойчивых выражений, которые в сущности потеряли свою семантическую наполненность. Самым частотным является «милости просим» (t:score — 19,85), обозначающее приглашение, когда кто-то просит присоединиться к какому-то мероприятию или участвовать в каком-то мероприятии. Оно имеет оттенок дружеского расположения в одних случаях или может носить шутливый непринужденный характер- в других, но если кто-то говорит неискренне, то может превратиться и в сарказм. Подобная мультиоттеночность свойственна и двум другим выражениям «скажите на милость» и «сделайте милость», где многое зависит от отношений между коммуникантами и от их намерений и настроения. Например, «по вашей милости»: «*По вашей милости мы получили эти два проекта». «А по вашей милости я*

теперь вообще 24 часа в сутки должна пахать? — набросилась телеведущая на главного техника телепроекта.

В результате исторических и социальных изменений, происходивших в российском обществе в XX веке, произошла трансформация компонентов значения понятия «милость». Свое первоначальное значение оно сохранило лишь в религиозной сфере, в то время как в светской культуре наблюдаются различные трансформации: от исчезновения духовного компонента, связанного с трансцендентным и этическим до полного стирания первоначального значения и зависимости от особенностей коммуникативной ситуации.

Литература

Фасмер М. Этимологический словарь русского языка. [Электронный ресурс]. URL: https://azbyka.ru/otechnik/ /Spravochniki/ etimologicheskij-slovar-russkogo-jazyka-fasmer/

Орлов М. прот. Понятие милости (Богословско-филологический опыт разграничения ἐλεεῖν и ἱλάσκεσθαι) [Электронный ресурс]. URL: azbyka.ru/otechnik/

Ринекер Ф., Майер Г. Библейская энциклопедия Брокгауза [Электронный ресурс]. URL: azbyka.ru/otechnik/Spravochniki/ biblejskaja-entsiklopedija-brokgauza/

Симфония по творениям святителя Григория Богослова [Электронный ресурс]. URL: https://azbyka.ru/otechnik/ Grigorij_Bogoslov/ simfonija-po-tvorenijam-svjatitelja-grigorija-bogoslova/

М. В. Хохлова
(Санкт-Петербург, Россия)
Санкт-Петербургский государственный университет
m.khokhlova@spbu.ru

ПОИСКОВЫЕ ИНТЕРФЕЙСЫ В КОРПУСАХ ТЕКСТАХ: СРАВНЕНИЕ ВОЗМОЖНОСТЕЙ И ОГРАНИЧЕНИЙ

Корпусы текстов претерпели значительные изменения за последние тридцать лет, последовательно развиваясь от простых инструментов, включающих тексты и поиск по ним, до усложненных систем, которые ориентированы на широкий круг пользователей. На примере НКРЯ продемонстрированы новые возможности, которые стали доступны в последней версии, и произведено сравнение с тем, что реализовано в иных корпусах (например, семейств Aranea и TenTen).

Ключевые слова: корпусы текстов, НКРЯ, поиск, интерфейс, русский язык

За последние тридцать лет корпусная лингвистика в России и в мире претерпела большие изменения. 1990-е и начало 2000-х ознаменовались созданием больших проектов по разработке корпусов текстов, объем которых превышал 100 млн слов и которые в дальнейшем получили название национальных. Следующий период был связан с автоматическим сбором данных из Интернета, идеей веба как корпуса (Web as Corpus), что привело к появлению сверхбольших корпусов, или гигакорпусов, включающих несколько десятков миллиардов слов. Периодизации в истории корпусной лингвистике посвящены работы [Захаров 2015; Солнышкина, Гатиятуллина 2020].

Распространение нейросетевых технологий, в частности, векторных представлений слов, трансформеров и больших языковых моделей, способствовало в некотором смысле пересмотру того, как должен выглядеть современный корпус. С одной стороны, поддерживается подход, зародившийся в корпусной лингвистике за рубежом, в рамках которого поисковые системы, используемые в корпусах, направлены на пользователя в широком смысле (им может быть как лингвист или лексикограф, так и иностранец, изучающий язык, учитель или школьник). Поэтому и сами инструменты, и выдаваемые результаты должны быть интуитивно понятными. С другой стороны, современные алгоритмы позволяют по-новому размечать текстовые данные больших объемов.

В нашей работе [Захаров, Бенко, Хохлова 2016] был дан обзор поисковых возможностей НКРЯ. В связи с появлением нового интерфейса, с августа 2022 ставшего окончательно доступным для поиска во всех корпусах в НКРЯ, возникла необходимость пересмотра тех технических опций, которые предоставляются пользователям. Ниже остановимся на некоторых из них.

Получение информации о сочетаемости слов с последующими квантитативными характеристиками стало доступным через функции коллокации, скетчи (в портрете слова) и частотность в лексико-грамматическом поиске [Савчук и др. 2024: 23–24]. Сравнение того, как представлены высокочастотные коллокации, извлеченные при помощи данных инструментов, дано в работе [Хохлова 2024]. В корпусе используются следующие широко распространенные метрики: t-score, MI3, LogDice и Loglikelihood. Дополнительно введена агрегированная мера (которая отсутствует в иных ресурсах). В корпусах семейств Aranea [Benko 2014] на платформе NoSketchEngine и TenTen [Jakubíček et al. 2013] на платформе Sketch Engine также применяется метрика MI (относящая к «классическим» при работе с автоматическим извлечением коллокаций). Несмотря на то, что она позволяет получить в верхней части отранжированного списка низкочастотные словосочетания, ее применение может быть обоснованным для нахождения редких сочетаний, терминов или ошибок в корпусе. В НКРЯ при поиске с помощью данных инструментов можно задавать не только грамматические, но и семантические и синтаксические признаки и отношения (последние — между ключевым словом и потенциальными коллокатами), что не предусмотрено в иных корпусах. Однако реализованные возможности в Aranea и TenTen позволяют задать минимальную частоту слов, образующих сочетание, а также находить коллокации, состоящие более чем из двух слов. Результаты не ограничены первой сотней примеров, что дает возможность получить более полное представление о сочетаемости тех или иных единиц. Под скетчами (или лексико-синтаксическими шаблонами) понимаются коллокации, упорядоченные по типу предварительно описанных синтаксических отношений [Хохлова 2021; Савчук и др. 2024]. Новая версия НКРЯ позволяет получить информацию о 10 наиболее частотных словосочетаниях, вместе с тем в других корпусах дополнительно осуществляется их кластеризация внутри синтаксических отношений.

Национальный корпус русского языка является ярчайшим примером того, как успешно может развиваться ресурс, при разработке которого используются подходы традиционной теоретической (и корпусной) лингвистики и новейшие технологические методы.

Литература

Захаров В. П. Корпуса русского языка // Труды Института русского языка им. В.В. Виноградова. 2015. № 6. С. 20–65.

Захаров В. П., Бенко В., Хохлова М. В. Кто ищет, тот найдет: поисковая функциональность корпусов русского языка // Прикладная лингвистика в науке и образовании. ALPAC REPORT — полвека после разгрома труды VIII Международной научной конференции, 2016. С. 158–163.

Савчук С. О., Архангельский Т. А., Бонч-Осмоловская А. А., Донина О. В., Кузнецова Ю. Н., Ляшевская О. Н., Орехов Б. В., Подрядчикова М. В. Национальный корпус русского языка 2.0: новые возможности и перспективы развития. // Вопросы языкознания, 2024, № 2. С. 7–34.

Солнышкина М. И., Гатиятуллина Г. М. История развития корпусной лингвистики (на примере англоязычных корпусов) // Вестн. Том. гос. ун-та. Филология. 2020. № 63. С. 132–160.

Хохлова М. В. Исследование лексико-синтаксической сочетаемости в русском языке с помощью статистических методов (на базе корпусов текстов): автореферат дис. ... кандидата филологических наук : 10.02.21 / Хохлова Мария Владимировна; [Место защиты: С.-Петерб. гос. ун-т]. СПб., 2011. 26 с.: с ил.

Хохлова М. В. Высокочастотные глагольные словарные коллокации с существительными в Национальном корпусе русского языка // Труды Института русского языка им В.В. Виноградова. 2024. № 4. С. 78–88.

Benko V. Aranea: Yet Another Family of (Comparable) Web Corpora. // Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, 2014. P. 257–264.

Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., Suchomel V. The TenTen corpus family. // 7th International Corpus Linguistics Conference CL, 2013. P. 125–127.

Л. Л. Шестакова, А. С. Кулева
(Москва, Россия)
Институт русского языка им. В. В. Виноградова РАН
lara.shestakova@mail.ru, an_kuleva@mail.ru

ИСПОЛЬЗОВАНИЕ РЕСУРСОВ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА В РАБОТЕ НАД «СЛОВАРЕМ ЯЗЫКА РУССКОЙ ПОЭЗИИ XX ВЕКА»

«Словарь языка русской поэзии XX века» представляет собой сводный объяснительный конкорданс (по творчеству десяти поэтов Серебряного века), содержащий всю лексику из выбранных источников. Начало реализации этого проекта относится к середине 1990-х гг., далее он осуществлялся параллельно с развитием корпусной лингвистики, с формированием Национального корпуса русского языка. В настоящее время обязательной частью работы составителей и редакторов словаря является обращение к корпусным ресурсам. Их использование позволяет более объективно описывать лексический массив словаря, особенности функционирования конкретной лексемы в поэтических текстах.

Ключевые слова: авторская лексикография, корпус, поэтический язык, словарь, словарная статья.

В середине 1990-х гг. в Институте русского языка им. В.В. Виноградова РАН под руководством В.П. Григорьева началась работа над «Словарем языка русской поэзии XX века» [СЯРП], нацеленным на описание языка десяти поэтов Серебряного века (см. [Шестакова 2022]). Еще в 1970-е гг. В.П. Григорьев видел будущее лексикографического представления языка поэзии в развитии компьютерных методов обработки текста, отслеживая новейшие на тот момент исследования [Григорьев 1973]. Технические возможности для составления масштабного сводного словаря появились много позже, благодаря достижениям Машинного фонда русского языка (ср. [Колодяжная 1987; Аношкина 1995]).

Начиная с раннего этапа работы над СЯРП составители столкнулись с рядом системных проблем. Обнаружилось, что оцифрованные тексты содержат ошибки сканирования и разметки, автоматический анализатор не справляется с задачами анализа поэтических текстов, а компьютерная обработка подготовленного составителями

материала создает новые сложности (связанные с потерей концов длинных строк, ударений, курсива и т.д.). Например, морфологический анализатор объединял в одну словарную статью не только омонимы, но и лексемы с «пересекающимися» грамматическими формами (*мутить*/*мутиться*/*мучить*, *тузить*/ *тужить*); неправильно восстанавливались начальные формы для редких или нестандартных лексем и имен собственных (*тамбовый*/*Тамбов*, *толстовец*/*толстовка*, *тьеполый*/*Тьеполо*, *убыточка*/*убыточек*, *умножарить*/ *умножарь*).

Со временем стало очевидно, что, вопреки запланированному, большу́ю часть работы необходимо делать в ручном режиме. В то же время возникшие в ходе работы трудности позволили более глубоко анализировать поэтический текст, его специфику, более тщательно составлять словарные статьи, уделяя, от тома к тому, всё большее внимание толково-комментирующей части словаря. Составители и редакторы СЯРП постоянно обращались как к словарям (общеязыковым и авторским), так и к исследовательской литературе, стараясь отражать интересные результаты изысканий в зонах значения и комментариев словарных статей. Важно, что с т. III научным консультантом СЯРП стал В.А. Плунгян, что нашло отражение в подаче случаев, трудных для словарного описания.

Новый этап работы над СЯРП начался с появлением [НКРЯ], и в особенности его Поэтического корпуса. В настоящее время обращение к корпусу является важной составной частью этой работы. Корпусные данные позволяют более объективно определять и описывать особенности функционирования конкретной лексемы: ее употребительность, стилистическую окраску, принадлежность к авторскому идиостилю и т.д. Отсутствие слова в НКРЯ или подтверждение единичности его употребления в произведении одного автора позволяет с уверенностью определять эту лексему как авторский неологизм. Наличие других употреблений дает возможность либо опровергнуть такую характеристику, либо — при явных отличиях в ее использовании — зафиксировать работу языковых механизмов, позволяющих создавать окказиональные единицы, а также проследить сознательное или неосознанное цитирование поэтов, их взаимовлияние. В последних опубликованных томах параллели такого рода фиксируются в зоне значения словарной статьи, например:

(1) ***ФОНАРЁК*** *[нов.?; ср. у А. Белого: …Близ бани, средь яблонь, у дома по грязи хлюпают ноги — и там, где плывет во тьме фонарек, да и там, где нет фонарька, хлюпают к бане ноги: это братья и сестры тянутся теперь на молитву («Серебряный голубь», 1909) и др.] Веди, веди, Егорка-/ Свет — карты поперек Родной! У Школьной Горки Пока что — ф. // Горит. — В чернильной, смольной Ночи́ — мечты игра: — Эх кабы вместо Школьной — Поклонная Гора! Цв928,29-38 (III,172)*

Работа над СЯРП как многолетним многотомным проектом близится к завершению: составители занимаются сейчас подготовкой к печати заключительного т. Х. Первоначально предполагалось, что завершенный словарь можно будет дополнять, расширяя круг поэтов. Однако на новом этапе СЯРП, содержащий систематизированное и выполненное по определенным правилам описание поэтического языка, может расширяться за счет обращения к НКРЯ. Кажется, что исследователи могут использовать оба источника в комплексе, поэтому содержанием следующего этапа работы над СЯРП будет его цифровое представление в полном объеме.

Корпус давно стал не только источником языкового материала: предоставив лингвисту богатый новый инструментарий, он задал и новый этап в исследовании языка [Плунгян 2008; 2024; Савчук 2019; 2022], что имеет большое значение и для цифровизации авторской лексикографии [Шестакова, Кулева 2023].

Литература

Аношкина Ж. Г. Подготовка частотных словарей и конкордансов на компьютере. М.: ИРЯ РАН, 1995. 60 с.

Григорьев В. П. (ред.). Поэт и слово. Опыт словаря. М.: Наука, 1973. 455 с.

Колодяжная Л. И. Автоматизированная лексикографическая система УНИЛЕКС. Словарно-ориентированная подсистема. М.: МГУ, 1987. 116 с.

НКРЯ — Национальный корпус русского языка [Электронный ресурс]. URL: http://ruscorpora.ru/

Плунгян В. А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. 2008. № 2(16). С. 7–20.

Плунгян В. А. Корпусная лингвистика на современном этапе // Вестник Российской академии наук. 2024. Т. 94, № 9. С. 773–780.

Савчук С. О. Корпус и изучение истории слов // Русская речь. 2019. № 2. С. 58–68. DOI:10.31857/S013161170004468-3

Савчук С. О. Мультимедийный поэтический корпус: общая структура и перспективы использования // Труды Института русского языка им. В. В. Виноградова. 2022. № 1 (31). С. 72–81. DOI: 10.31912/pvrli-2022.1.8

СЯРП — Словарь языка русской поэзии XX века. Т. I–IX– / Григорьев В.П. (отв. ред.), Шестакова Л.Л. (отв. ред.), Колодяжная Л.И. (ред.), Кулева А.С. (ред.), Бакеркина В.В., Гик А.В., Реутт Т.Е., Фатеева Н.А. М.: Языки славянской культуры, 2001–2022–.

Шестакова Л. Л. «Словарь языка русской поэзии XX века»: От пробного выпуска до VIII тома // Труды Института русского языка им. В.В. Виноградова. 2022. № 2 (32). С. 148–164. DOI 10.31912/pvrli-2022.2.10

Шестакова Л. Л., Кулева А. С. Авторская лексикография в электронно-цифровую эпоху // Terra Linguistica. 2023. Т. 14, № 3. С. 94–113. DOI: 10.18721/JHSS.14308

К.М. Шилихина
(Воронеж, Россия)
Воронежский государственный университет
shilikhina@gmail.com

КОРПУСНЫЕ ДАННЫЕ В ИССЛЕДОВАНИЯХ МЕТАКОММУНИКАЦИИ

В докладе обсуждаются возможности использования языковых корпусов для изучения метакоммуникативных комментариев лексического выбора. Такие комментарии эксплицируют рефлексию говорящего относительно выбранных языковых средств, фокусируют внимание адресата на определенных способах номинации и дают возможность имплицитно передавать дополнительные прагматические смыслы. В исследовании используются данные различных подкорпусов Национального корпуса русского языка, позволяющие дать количественную и качественную интерпретацию данных об использовании метакоммуникативных комментариев в различных модусах дискурса.

Ключевые слова: языковой корпус, метакоммуникация, метапрагматика, рефлексия говорящего, конструкция, лексический выбор.

В современной лингвистике достаточно много внимания уделяется метакоммуникации, поскольку метакоммуникативная часть высказывания отражает процесс регулирования дискурса [Hyland 2005] и социальных отношений между его участниками [Чернявская 2020]. Обращаясь к теме метакоммуникации, исследователи выделяют ряд проблем. Наиболее очевидной является проблема терминологии — на сегодняшний день границы понятий, с помощью которых описывается «*мета-*» часть функционирования языка, не установлены, и поэтому результаты исследований во многом зависят от того, что понимается под терминами *метаязык, метадискурс, метапрагматика, метакоммуникация*. В ряде исследований эти термины используются как синонимы, однако, например, в работе [Hyland 2017] предлагается разделение *метадискурса* и *метапрагматики*: если первым термином обозначаются языковые средства, обеспечивающие взаимодействие говорящего и адресата и регулирующие процесс понимания, то второй связан с анализом языковых средств, с помощью которых мы демонстрируем осознанное отношение к использованию языка.

Представляется, что наиболее общим понятием, которое включает в себя и метапрагматические, и метадискурсивные элементы, является *метакоммуникация*. И, поскольку в данном исследовании описываются конструкции, семантика которых направлена и на язык (оценку языковых средств), и на участников дискурса (маркирование лексического выбора привносит в высказывание дополнительные прагматические значения), мы будем говорить именно о *метакоммуникации*.

Вторая проблема, с которой сталкиваются исследователи метакоммуникации, — это проблема источников данных и эмпирического материала. В англоязычных работах материалом исследований чаще всего выступает письменный академический дискурс [Hyland 2005] либо электронная академическая коммуникация [Metadiscourse in Digital Communication 2021]. В результате то, что на сегодняшний день известно о метакоммуникации, в значительной степени основано на письменной англоязычном академическом дискурсе. Кроме того, у исследователей возникают вопросы относительно объема используемых данных: многие исследования представляют собой анализ отдельных источников (статей, блогов или выступлений), что не позволяет делать более широкие обобщения.

Обращение к корпусным данным кажется естественным способом решения проблемы данных. Однако некоторые исследователи, например [Ädel & Mauranen 2010] критически оценивают возможности применения корпусов в исследованиях метакоммуникации и отдают предпочтение квалитативным исследованиям конкретных текстов. Причиной такого отношения является то, что языковые единицы, которые используются как метакоммуникативные маркеры, часто могут выполнять и другие функции, поэтому необходимо вручную отбирать только те контексты, в которых реализуется именно метакоммуникативная функция.

Даже в тех случаях, когда лингвисты обращаются к корпусам, как правило, они ограничиваются изучением «однословных» метакоммуникативных маркеров, например, перформативных глаголов, слов, оценивающих или называющих речевые действия (шутить, иронизировать). Однако метакоммуникация осуществляется и с помощью конструкций (например, *как сейчас принято говорить, если можно так сказать, что называется*). Такие конструкции

эксплицитно отражают рефлексию говорящего относительно языка и позволяют делать акцент на том, *как* говорящий использует язык.

Поскольку для описания того, как и с какими целями говорящие прибегают к метакоммуникации, необходим не только качественный, но и количественный анализ, очевидный «плюс» в использовании корпусов для изучения конструкций, выполняющих метакоммуникативную функцию, заключается в возможности получать количественные данные и делать выводы относительно того, насколько метакоммуникация распространена в различных типах и модусах дискурса. Основной, газетный, устный подкорпусы, а также подкорпус «Социальные сети» НКРЯ дают возможность сопоставлять частотность использования различных метадискурсивных конструкций в разных типах дискурса. Однако здесь возникает потребность в «ручном» анализе конкорданса, поскольку интересующие нас конструкции могут употребляться и не как метапрагматические комментарии:

[1] *Это, **я бы сказал**, псевдоисторическая литература, не научная, а порожденная информационно-психологической войной, фактически не прекращающейся во всем мире. [В. В. Седов. Этногенез ранних славян // «Вестник РАН», 2003]*

[2] ***Я бы сказал**, что это один из самых амбициозных планов реформы, и отступать от него мы не намерены. [Александр Чудодеев. Расписание на завтра // «Итоги», 2003.02.04]*

Вообще, соответствие формы и функции конструкции можно отнести к одному из наиболее спорных вопросов в использовании корпусов в изучении метакоммуникации. Процесс ручного исключения нерелевантных примеров необходим для того, чтобы избежать поверхностных предположений о соответствии формы и функции.

Количественный анализ, проведенный после ручной обработки конкорданса, позволяет делать выводы о частотности метакоммуникативных конструкций в различных сферах дискурса: самый большой объем метакоммуникации характерен для текстов СМИ, за ними следуют социальные сети и устная речь. Качественный анализ конкордансов позволяет выделять дополнительные прагматические смыслы, которые вносят в высказывание метакоммуникативные языковые средства, например, «осовременивание» номинации,

маркирование более престижного способа номинации, самопрезентация говорящего и др.

Таким образом, корпусное изучение метакоммуникативного использования языка позволяет делать обобщения, касающиеся метакоммуникации в различных модусах дискурса, однако требует сочетания статистической и ручной обработки для получения «чистых» количественных данных и дальнейшей интерпретации прагматических смыслов, которые демонстрируют прагматическую вариативность в функционировании метакоммуникативных языковых средств.

Литература

Чернявская В.Е. Метапрагматика коммуникации: когда автор приносит свое значение, а адресат свой контекст // Вестник СПбГУ. Язык и литература. 2020. Т.17. Вып.1. С. 135-146.

Ädel A., Mauranen A. Metadiscourse: Diverse and divided perspectives // Nordic Journal of English Studies. 2010. № 9 (2). P. 1-11.

Hyland K. Metadiscourse: Exploring interaction in writing. Continuum, 2005. 2 p.

Hyland K. Metadiscourse: What is it and where is it going? // Journal of Pragmatics. 2017. № 113. P. 16-29.

Metadiscourse in Digital Communication. New Research, Approaches and Methodologies. Ed. by L. D'Angelo, A. Mauranen & S. Maci. Cham: Palgrave Macmillan, 2021. 169 p.

Р. И. Шмурак
(Ханчжоу, Китай)
Институт русского языка и культуры,
Школа международных исследований, Чжэцзянский университет;
roman.shmurak@gmail.com

НЕЙРОННЫЕ СЕТИ КАК ИНСТРУМЕНТ КОРПУСНОГО ПОИСКА

Доклад посвящен перспективам использования ChatGPT и других нейросетей в работе с корпусами. Несмотря на как минимум 2-х летнюю историю ChatGPT и его аналогов, взаимодействия нейросетей и корпусов в настоящее время недостаточно полно представлены в лингвистических исследованиях. Взаимодействие это понимается, прежде всего, с технической позиции, а также в аспекте обучения нейросетей языковому взаимодействию с пользователем. При этом вне рассмотрения остается потенциал нейросетей в качестве инструмента корпусного поиска. В фокусе внимания находится методологическая составляющая вопроса.

Ключевые слова: ChatGPT, НКРЯ, нейросеть, корпус, перспективы

Когда речь заходит об использовании нейросетей в лингвистике, первые ассоциации представлены чаще всего двумя парадигмами:

1. техническая сторона вопроса — создание алгоритмов использования, а также «софта» (англ. «software» — программного обеспечения);
2. обучение нейросетей языковому взаимодействию с пользователем.

Возникновение обеих ассоциаций ожидаемо и закономерно. В этой связи совершенно неудивительно взрывное число работ, посвященных проблемам перевода с помощью нейросетей (см. например [Гуров 2021, Осипов 2023, Ширяева 2024]). Немало можно найти работ по взаимодействию пользователя-лингвиста и нейросети как его инструмента (См. например [Бейненсон 2024, Косых 2022а, Мамонтов 2024]). И это только работы лингвистического профиля, исследований нейросетей с позиции IT, разумеется, насчитывается в разы больше (См. например [Косых 2022б, Чучупал 2020]).

На этом фоне вызывает некоторое удивление малое количество

работ, посвященных взаимодействию нейросетей и корпуса. При базовом поиске по ключевым словам в eLibrary просто не удалось обнаружить исследований, которые рассматривали бы взаимодействие корпуса и нейросети не в парадигме «корпус как инструмент для нейросети», а в реверсивной парадигме «нейросеть как инструмент для корпуса». Это довольно странно, поскольку с самого зарождения корпусов перед лингвистом стояло три проблемы, которые ставили под удар самое главное достижение корпусной лингвистики — ее доказательность и объективность:

1. «Первый шаг», проблема субъективного корпусного запроса на основе интуиции;

2. «Последний шаг», проблема субъективной интерпретации полученных с помощью корпуса данных;

3. «Кризис запроса», отсутствие выдачи на корпусный запрос не всегда говорит о провале гипотезы, но может свидетельствовать о несовершенстве самого запроса.

На наш взгляд, использование нейросетей в их текущем состоянии способно повысить объективность корпусного поиска, а в будущем, вероятно, довести ее полностью до 100%. Ниже на базе ChatGPT и НКРЯ представлен вариант подобной работы с нейросетями и корпусами.

< Примеры >

Алгоритм эксперимента:

1. Генерирование запроса в нейросеть на выдачу фразеологических единиц, способных в зависимости от контекста иметь положительную или отрицательную коннотацию;

2. Ручной отбор данных, предложенных нейросетью;

3. Ввод отобранных данных в НКРЯ;

4. Сортировка на 3 вида всех примеров выдачи в НКРЯ по одной фразеологической единице (например, «*смотреть сквозь пальцы*»):

— с очевидной положительной коннотацией;

— с очевидной отрицательной коннотацией;

— с неясной коннотацией;

5. Данные, полученные путем корпусного поиска и имеющие неясную тональность, вводятся в нейросеть с генерированием запроса на их сентимент-анализ. Одновременно с нейросетью и независимо от нее, исследователь также выполняет сентимент-анализ данных;

6. Сравнение и интерпретация результатов исследователя и ChatGPT, выводы.

Результаты сентимент-анализа, произведенного ChatGPT и исследователем, далеко не всегда имеют совпадение. Независимо от того, согласится ли исследователь с выводами и аргументами нейросети или продолжит эксперимент с определением тональности, такое разночтение в любом случае позволяет сделать два вывода:

Во-первых, расхождение интерпретаций ChatGPT с мнением исследователя недвусмысленно иллюстрирует всю серьезность проблемы «последнего шага» и насколько велика вероятность ошибки при трактовке корпусных данных.

Во-вторых, представленная работа с нейросетью и корпусом хорошо демонстрирует возможности нейросети в решении проблемы «последнего шага» и свидетельствует о значительном потенциале интеграции корпуса и нейросети в исследовательской деятельности.

Литература

Бейненсон В. А. К вопросу о субъектности генеративных нейросетей: соавтор или инструмент? / В. А. Бейненсон // MEDIAОбразование. Цифровая среда: между позитивом и деструкцией : Сборник материалов VIII Международной научно-практической конференции. Челябинск: Челябинский институт развития профессионального образования. 2024. С. 170–173.

Гуров А. Н. Художественный перевод как непреодолимое препятствие для нейросетей / А. Н. Гуров // Казанская наука. 2021. № 7. С. 81–83.

Косых Н. Е. Разработка WEB-приложения для анализа настроений текста с помощью фреймворка Flask и языка Python / Н. Е. Косых, А. Д. Хомоненко, О. Н. Куранова // Наукоемкие технологии в космических исследованиях Земли. 2022а. Т. 14, № 1. С. 45–52.

Косых Н. Е. Особенности предварительной обработки текста для проведения анализа настроений / Н. Е. Косых, И. А. Молодкин, А. Д. Хомоненко // Интеллектуальные технологии на транспорте. 2022б. № 3(31). С. 68–73.

Мамонтов А. И. Преобразование тональности текста с помощью нейронных сетей / А. И. Мамонтов, Б. Н. Нургатин. 2024. С. 3–7.

Осипов Д. В. Языковые барьеры в метавселенных: сила нейронных сетей в переводе / Д. В. Осипов // Евразийский филологический вестник. 2023. №. 2(2). С. 21–39.

Чучупал В. Я. Нейросетевые модели языка для систем распознавания речи / В. Я. Чучупал // Речевые технологии. 2020. № 1–2. С. 27–47.

Ширяева А. А. Перспективы использования машинного перевода на основе нейросетей при переводе текстов официально-делового стиля / А. А. Ширяева Д. Ю. Леонова // Философия и наука в культурах Запада и Востока : Сборник статей по материалам VII Всероссийской научной конференции с международным участием. Томск: Национальный исследовательский Томский государственный университет. 2024. С. 146–150.

К.А. Щукина
(Санкт-Петербург, Россия)
СПбГУ
k.shukina@spbu.ru

К ВОПРОСУ ОБ ИСПОЛЬЗОВАНИИ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА В ВЫПУСКНЫХ КВАЛИФИКАЦИОННЫХ РАБОТАХ

Статья посвящена рассмотрению использования функционала Национального корпуса русского языка в выпускных квалификационных работах студентов СПбГУ, а также сравнению корпусных данных и данных словарей русского языка; кроме того, рассматриваются возможности использования моделей дистрибутивной семантики.

Ключевые слова: дистрибутивная семантика, Национальный корпус русского языка, RusVectōrēs, семантические отношения

Последние годы на кафедре русского языка как иностранного Санкт-Петербургского государственного университета появляется значительное количество бакалаврских и магистерских работ, выполненных в рамках функционального подхода. Возьмем для примера темы работ за последние несколько лет, посвященные синонимам и синонимическим рядам: «Синонимический ряд глаголов с доминантой «исчезнуть»: функционально-семантический аспект (на фоне корейского языка)», «Семантика и функционирование глагольных синонимов с доминантой «бояться» (на фоне китайского языка)» и др. Материалом для подобного рода работ служат данные словарей синонимов и толковых словарей русского языка, а также материалы сайта «Национальный корпус русского языка». Традиционно по словарям определяется семантика, а по материалам корпуса выявляются особенности функционирования лексических единиц.

Нам представляется интересным некоторое расширение спектра работ в область дистрибутивной семантики, возможности которой явно недооцениваются в лингвистических исследованиях. Этому вопросу, в

частности, посвящены две работы М.К. Тимофеевой: «Возможности использования сервиса RusVectōrēs для выявления семантических ассоциатов глаголов русского языка» и «Типология семантических отношений, выявляемых посредством инструмента RusVectōrēs» [Тимофеева, 2018а, 2018б], где автор предлагает классификацию семантических отношений и подчеркивает, что «сервис RusVectōrēs выявляет довольно большое количество синонимов, однако сравнение с данными словаря синонимов русского языка (…) показывает, что доля обнаруживаемых словарных синонимов не очень велика (17,61 %)» [Тимофеева, 2018а], а также отмечает, что «при идентификации семантической связи между входным словом и словом, выявленным посредством RusVectōrēs, рассматриваются следующие возможности: выявленное слово может быть по отношению к заданному слову синонимом, антонимом, гипонимом, гиперонимом, холонимом, меронимом, признаком, операцией, ситуационно связанным понятием, словообразовательным вариантом» [Тимофеева, 2018б].

Возвращаясь к работам, выполненным на кафедре русского языка как иностранного СПбГУ, обратимся к выпускной квалификационной работе бакалавра Ю.В. Володы «Синонимический ряд прилагательных с доминантой «вежливый»: лексико-семантический, лингвокультурологический и лексикографический аспекты» [Волода, 2021]. В своей работе Ю.В. Волода анализирует данные словарей синонимов и на их основе составляет сводный синонимический ряд с доминантой «вежливый» из прилагательных, входящих в 3 и более словаря: *вежливый* (5), *учтивый* (5), *любезный* (5), *обходительный* (5), *корректный* (4), *деликатный* (4), *галантный* (3), *предупредительный* (3) (там же) (см. Таблица 1).

Данные, полученные из Ю.В. Володой, мы сравнили с данными сервисов «Похожие слова» Национального корпуса русского языка и проекта-спутника RusVectōrēs, в которых отображаются ближайшие семантические ассоциаты выбранного слова. В первом столбце таблицы мы приводим данные из НКРЯ, ранжировав их по значению коэффициента семантической близости.

Таблица 1. Синонимический ряд с доминантой «вежливый» в сравнении с корпусными данными

«Похожие слова» (НКРЯ)	**«Похожие слова» (RusVectōrēs по НКРЯ)**	**Данные Ю.В. Володы**
учтивый (0.884902)	*учтивый* 0.69	*учтивый* (5)
обходительный (0.734349)	*обходительный* 0.61	*обходительный* (5)
тактичный (0.714415)	*тактичный* 0.64	*тактичный* (Гаврилова)[1]
приветливый (0.709105)	*приветливый* 0.61	*приветливый* (Гаврилова)
предупредительный (0.708281)	*предупредительный* 0.61	*предупредительный* (3)
-	деликатный 0.61	деликатный (4),
дружелюбный (0,70118)	*дружелюбный* 0.60	отсутствует в словарях синонимов
доброжелательный (0,684435)	*доброжелательный* 0.59	отсутствует в словарях синонимов
-	*корректный* 0.59	корректный (4)
-	*сдержанный* 0.59	отсутствует в словарях синонимов
-	-	*любезный* (5)
-	-	*галантный* (3)

Из таблицы видно, что данные словарей лишь отчасти совпадают с корпусными данными, так, в частности, лексемы «тактичный» и «приветливый» встречаются лишь в Словаре синонимов и антонимов А.Н. Гавриловой, лексема «деликатный» есть в 4 словарях синонимов,

[1] Гаврилова — здесь и далее: Гаврилова А.С. Словарь синонимов и антонимов современного русского языка. 50000 слов / Под ред. А.С. Гавриловой // М.:«Аделант», 2014. – 800 с.

но НКРЯ не показывает ее семантическую близость со словом «вежливый». С лексемами «любезный» и «галантный», на наш взгляд, ситуация немного другая, поскольку для «галантный» словари выдают помету «устаревший»; для лексемы «любезный» такой пометы нет, но, с нашей точки зрения, в современном русском языке она постепенно переходит в разряд устаревших, что и подтверждается корпусными данными.

Таким образом, при анализе данных словарей и сопоставлении их с корпусными данными мы можем получить довольно интересные результаты, отражающие состояние современного русского языка не только в отношении прилагательных, взятых нами для примера, но и для других частей речи.

Литература

Володa Ю.В. Синонимический ряд прилагательных с доминантой «вежливый»: лексико-семантический, лингвокультурологический и лексикографический аспекты. Научный руководитель: Е.И. Зиновьева. СПбГУ, 2021 URL: https://dspace.spbu.ru/bitstream/11701/29971/1/VKR_Voloda_U.V.docx (дата обращения: 07.07.2024)

Тимофеева М.К. Возможности использования сервиса RusVectōrēs для выявления семантических ассоциатов глаголов русского языка // Научный диалог. 2018. №9. URL: https://cyberleninka.ru/article/n/vozmozhnosti-ispolzovaniya-servisa-rusvect-r-s-dlya-vyyavleniya-semanticheskih-assotsiatov-glagolov-russkogo-yazyka (дата обращения: 07.07.2024) — 2018а

Тимофеева М.К. Типология семантических отношений, выявляемых посредством инструмента RusVectōrēs // Научный диалог. 2018. №8. URL: https://cyberleninka.ru/article/n/tipologiya-semanticheskih-otnosheniy-vyyavlyaemyh-posredstvom-instrumenta-rusvect-r-s (дата обращения: 07.07.2024) — 2018б

Международная научная конференция,

посвященная 20-летию Национального корпуса русского языка

Материалы конференции

Подписано в печать 20.12.2024 г. Формат 60x84/16.
Бумага офсетная. Печ. л. 12. Тираж 200 экз.

Федеральное государственное бюджетное учреждение науки
Институт русского языка им. В.В. Виноградова
119019, г. Москва, ул. Волхонка, д. 18/2