Тезисы
всероссийской конференции
«От языковых машинных фондов к лингвистическим корпусам: памяти В.М. Андрющенко»

Лаборатория автоматизированных лексикографических систем
НИВЦ МГУ имени М.В. Ломоносова
Институт русского языка имени В.В. Виноградова

Москва, 28 – 29 сентября 2018 г.

От составителей

Необходимость построения языковых машинных фондов – компьютерных хранилищ лингвистических данных – была осознана в конце 70 х – начале 80-х гг. прошлого века. Теоретическим проблемам «машинного фонда данных для автоматизированной системы лексикографических исследований» были посвящены три всесоюзных конференции, проводившихся в 1983, 1987 и 1989 гг. В 1985 г. по инициативе академика А.П. Ершова начались работы по созданию машинного фонда русского языка. В работе над Машинным фондом принимало участие более 40 организаций-соисполнителей; ведущей организацией стал Институт русского языка им. В.В. Виноградова, а Лаборатория вычислительной лингвистики (ныне Лаборатория автоматизированных лексикографических систем) Научно-исследовательского вычислительного центра МГУ под руководством В.М. Андрющенко превратилась в одну из основных площадок проекта: именно здесь вводом в компьютер и лемматизицией Саратовского корпуса было положено начало русского корпуса устной речи, именно здесь был оцифрован Грамматический словарь А.А. Зализняка, без которого сейчас не мыслима никакая автоматическая обработка русского текста. Тогда же параллельно с работой над компонентами машинного фонда русского языка началась работа над машинными фондами языков народов СССР.

На заре создания машинных фондов вряд ли кто-то ожидал, что компьютерная техника, на которую в то время смотрели с пиететом и некоторым недоверием, будет развиваться столь стремительно, что разнообразные ухищрения с целью экономии компьютерной памяти или кодирования нестандартной графики станут неактуальными, электронные текстовые корпусы достигнут невообразимых объемов, а их автоматическая обработка – немыслимой скорости, и только ручное индексирование останется узким местом на пути к светлому лингвистическому будущему. Сегодня вместо текстового модуля Машинного фонда русского языка у нас есть Русский национальный корпус, вместо компьютерных словарей – электронные словари онлайн. Интернет, еще одна новая технология, появившаяся тех пор, сделала мир безграничным, а корпусы и словари общедоступными.

Задача конференции – обсудить развитие тех идей, которые легли в основу языковых машинных фондов, того, в какой степени эти идеи

реализованы в современных корпусах, лексических базах и лингвистических программных средствах, что было приобретено и что, возможно, потеряно с развитием технических средств и в какой степени современные корпусы и словарные базы могут считаться наследниками первых модулей машинных фондов.

Тезисы публикуются в авторской редакции.

Использование математики в лингвистике

Алпатов Владимир Михайлович
Институт языкознания РАН
(Москва)

До середины XX в. лингвистика и математика развивались независимо друг от друга. Однако уже И.А. Бодуэн де Куртенэ писал о желательности введения математики в лингвистическое образование и предсказывал ее использование в науке о языке в будущем, а Е.Д. Поливанов отмечал необходимость для нее статистических методов. На первом этапе развития структурализма на идеи его представителей оказывала влияние математическая логика, но собственно математический аппарат не применялся.

Математическая лингвистика получила развитие уже после второй мировой войны, первоначально в рамках структурализма. Она начала формироваться сначала на Западе, прежде всего, в США, а во второй половине XX в. и в СССР. После формирования генеративной лингвистики построение математических моделей и широкое применение математического аппарата стали выглядеть необходимой частью новой науки о языке.

В СССР в 1960-1970-е гг. считавшиеся передовыми лингвистические направления активно использовали математические методы. В это время под влиянием значительных научно-технических достижений и в связи с активизацией прикладных исследований считалось, что уровень развития той или иной науки определяется степенью математизации. Гуманитарные науки оценивались как «отсталые», но лингвистику признавали наименее «отсталой» среди них, поскольку математику там активно использовали. Полная математизация науки о языке казалась задачей близкого будущего.

И.А. Мельчук писал, что контрольным критерием при проверке построений лингвиста должна быть «принципиальная осуществимость модели или любого ее фрагмента на вычислительной машине» [Мельчук 1974: 20].

Использование математики в лингвистике в основном сводилось к двум не связанным между собой областям: математической логике и математической статистике. Из них первостепенное внимание уделялось математической логике, на основе методов которой строились формальные модели языка, тогда как статистика имела вспомогательное значение.

В СССР не все лингвисты были сторонниками математизации. Существовала и «традиционная» лингвистика, представители которой допускали использование статистики, но возражали против логических моделей. Особенно резким было выступление В.И. Абаева [Абаев 1965], которого его противники обвинили в «мракобесии» и невежестве.

Построение логико-математических моделей в так называемой формальной лингвистике (прежде всего, генеративистами, но не только ими) происходит и сейчас. Продолжаются традиции, заложенные структуралистами, хотя основное внимание вместо фонологии и морфологии уделяется синтаксису. Однако в России сейчас всё более преобладает функциональная лингвистика, которая не ограничивается изучением устройства языка и ставит задачу исследования его функционирования. И здесь уровень математизации (исключая, правда, прикладные исследования) упал почти до нуля. Сравнительно молодой лингвист А.Ч. Пиперски недавно написал, что применение математики в других науках, включая лингвистику, сводится, прежде всего, к статистическим методам, а логический анализ имеет лишь узкую сферу применения [Пиперски 2017: 136]. То, что он пишет, очень похоже на то, что полвека назад высказывал В.И. Абаев.

Вероятно, отказ (временный?) от математизации связан с большей сложностью объекта исследования функциональной лингвистики. Пока неясно, как можно здесь что-то формализовать. А вот статистика всегда полезна. Сейчас в лингвистике популярны опросы и анкетирование, но, к сожалению, при этом чаще всего исследователи методами математической статистики элементарно не владеют.

ЛИТЕРАТУРА

Абаев В.И. Лингвистический модернизм как дегуманизация науки о языке // *Абаев В.И.* Статьи по теории и истории языкознания. М.:, Наука, 2006 (1 издание – 1965). С. 108-131.

Мельчук И.А. Опыт теории лингвистических моделей «Смысл ↔ Текст». Семантика, синтаксис. М.: Наука, 1974.

Пиперски А.Ч. Рецензия на: *N. Levshina.* How to do linguistics with R. Amsterdam: John Benjamins, 2015 // Вопросы языкознания, 2017, №2. С. 134-139.

Маркемный анализ как разновидность компьютерного анализа текста

Артемова Ольга Григорьевна
ВГТУ
(Воронеж)

Изучению языка писателя посвящено множество работ, но до сих пор нет четких параметров, позволяющих объективировать не только характеристику стиля автора, но и проследить взаимосвязь разных авторов, а также исследовать формирование языка литературы. В связи с этим не прекращаются попытки решить задачу объективации и формализации содержательного анализа текста, в первую очередь – литературного.

Маркемология – молодое направление в лингвистике, развиваемое А.А. Кретовым и его учениками, позволяет, используя метод маркемного анализа, формализовать содержательный анализ текстов [Кретов, 2007; 2010]. Маркемный анализ – это метод компьютерного выделения ключевых слов – маркем. Маркема – это одна из 50 прошедших через все фильтры словоформ с максимальным положительным значением индекса текстуальной маркированности словоформ (ИнТеМ). В качестве потенциальных маркем рассматриваются имена существительные, как наименее маркированная и ориентированная на внеязыковую действительность часть речи, которые прошли специальную систему фильтров, позволяющих исключить попадание в разряд маркем

неинформативной лексики. Систему фильтров образуют 1) частеречный, 2) грамматический, 3) грамматико-семантический, 4) тематико-семантический, 5) стилистический, 6) диалогический и 7) классификационный фильтры. Их характеристика подробно изложена в статье А.А. Фаустова и А.А. Кретова «Понятие маркемы и предварительные итоги маркемного анализа русской литературы» [Фаустов, 2017]. Целесообразным считается выделение 50 маркем для одного автора.

Используемый для выделения маркем базовый параметр ИнТеМ выражает зависимость между длиной словоформы и частотой ее употребления и представляет собой авторскую составляющую, обусловливающую частоту употребления словоформы в тексте.

Математически ИнТеМ представляет собой разность, полученную в результате вычитания веса словоформы по длине из ее веса по частоте. С позиции лингвистики – это степень субъективной, т.е. текстовой, весомости отдельной словоформы для конкретного текста. Вычисление ИнТеМа английских словоформ осуществляется программным комплексом тематического анализа лексики “ProTemAL-Engl” (автор – А.С. Гусельникова, научный руководитель – д. техн. наук. И.Е. Воронина, научный консультант – д.ф.н. А.А. Кретов) посредством обработки собранных в единый файл текстов. Для вычисления ИнТеМа русских словоформ используется программа «ТемАЛ» (автор – Ирина Попова, научный руководитель - д. техн. наук. И.Е. Воронина, научный консультант – д. филол.н. А.А. Кретов).

При выделении маркем основное внимание обращается на словоформы, которые можно считать наиболее значимыми для автора текста. К ним относятся: абстрактная (безденотатная) лексика, обозначающая важнейшие категории культуры, описывающая внутренний мир человека и его взаимоотношения с социумом; базовые понятия, соответствующие аспектам универсальных философских концептов; лексемы, обозначающие природные объекты и т.д. Из артефактов – только употребляющиеся в символическом значении. Таким образом, признание словоформы маркемой определяется ее семантикой, а в спорных случаях – обращением к контексту.

При проведении сопоставительного маркемного анализа по ряду объективных причин величину ИнТеМа не всегда можно считать достоверным показателем. В таких случаях целесообразно использовать Нормированный ИнТеМ, представляющий собой результат деления ИнТеМа

каждой маркемы автора на Суммарный ИнТеМ всех его маркем. Тогда значение нормированного ИнТеМа всегда располагается в интервале 0 – 1, что позволяет максимально объективировать сопоставление маркемных наборов всех исследуемых авторов.

Сопоставительный маркемный анализ предполагает составление сводного списка маркем, составляемого из маркемных списков исследуемых авторов и ранжированного в порядке убывания суммарного нормированного интегрального веса его маркем.

Величина суммарного нормированного интегрального веса определяется как произведение суммарного нормированного ИнТема на количество авторов, употребивших данную маркему. Первые 50 маркем ранжированного сводного списка являются маркемами, подлежащими анализу. Это могут быть маркемы среза, периода или общие маркемы отдельных авторов.

Для установления литературной близости писателей применяется коэффициент корреляции, вычисляемый в каждой паре авторов на основе их общих маркем. Его величина равна произведению суммарных нормированных ИнТеМов общих маркем двух авторов.

Итак, применяемый нами метод маркемного анализа позволяет решать комплекс задач, стоящих перед маркемологией – исследование творчества отдельных авторов; составление общесрезовых (односрезовых) или общеинтервальных (многосрезовых) маркемных списков; описание специфики маркем отдельных авторов или групп авторов; изучение эволюции маркемной лексики на основе исследования динамики маркем в нескольких хронологических срезах; влияние социокультурных процессов на динамику маркем. Маркемологические исследования также позволяют посредством выделения общих маркем и определения на их основе коэффициента корреляции авторов устанавливать взаимосвязи различных авторов и таким образом обнаруживать преемственность в литературе [Кретов, 2009; Кашкина, 2013] .

ЛИТЕРАТУРА

1. Кретов А.А. Метод формального выделения тематически нейтральной лексики (на примере старославянских текстов) / А.А. Кретов //

Вестник ВГУ. Серия Системный анализ и информационные технологии, 2007, № 1, С.81 – 90.

2. Кретов А.А. Понятие маркемы: методика выявления и практика использования / А.А. Кретов // Универсалии русской литературы: сб-к статей. – Воронеж, 2010. – С. 138 – 153.

3. Фаустов А.А. Понятие маркемы и предварительные итоги маркемного анализа русской литературы / А.А. Фаустов, А.А. Кретов // Вестник Воронежского гос. ун-та. Сер. Лингвистика и межкультурная коммуникация. 2017, № 4. С. 16 – 31.

4. Кретов А.А. Архаисты и новаторы в русской литературе XVIII – начала XX вв. / А.А. Кретов // Универсалии русской литературы: сб-к статей. – Воронеж: Воронежский государственный университет; Издательский дом Алейниковых, 2009, С.29 – 48.

5. Кашкина А.А. Маркемный анализ языка русской поэзии: автореферат дис. …канд. филол. наук. – Воронеж, 2013. – 25 с.

О способе обработки естественно-языковых текстовых данных с автоматическим построением семантической сети

Бочаров Николай Викторович
DWORQCOM
(Москва)

При обработке естественно-языковых текстовых данных исследователи прежде всего сталкиваются с проблемами, напрямую связанные как с морфологической, так и с синтаксической омонимией слов, и выражающиеся в трудности распознавания плавающих признаков принадлежности одних и тех же слов к разным членам предложения.

К морфологической омонимии относятся слова, имеющие одно написание, но отличающиеся или значением, или ударением. Отдельно стоит проблема, вызванная небрежностью обращения с буквой «ё».

К синтаксической омонимии относится как проявление принадлежности одного и того же слова к разным частям речи, так и проявление присутствия в одних и тех же словах плавающего ударения. Сюда же относится проблема с порядковыми и количественными числительными,

идентифицирование семантики которых осуществляется в тексте принципиально по разным алгоритмам. Отдельно стоит «синтетическая» синтаксическая омонимия, проявляющаяся в исключении той или иной части речи у отдельно взятого слова, несмотря на присутствие в этом слове её признака.

Следует отметить, что омонимия имеет разную природу своего происхождения. Графематические корни выявляются: с числительными – в случае написания их цифрами; с именами собственными – в случае написания омонимов с прописной буквы; с омофонами – в случае неправильной разбивки звукоряда на слова в системах распознавания речи. Морфологические корни выявляются в омоформах и в омографах. Отдельно следует выделить проблему присутствия омонимов в предложных группах, когда разрешение омонимии приводит к различным синтаксическим ролям всей предложной группы в предложении.

Природа происхождения перечисленных выше случаев омонимии указывает на моменты их обнаружения на соответствующих этапах лингвистического анализа. При этом у разработчика семантического анализатора текста возникает проблема выбора этапа для разрешения омонимии, на котором принимается решение об ограничении количества морфологических признаков у лексемы-омонима. В большинстве случаев таким моментом является этап синтаксического анализа текста. Учитывая тот факт, что разрешённая омонимия может повлиять на различные признаки лексемы, вплоть до графематических, будет справедливым говорить о применении технологии запараллеливания этапов семантического анализа текста.

Отталкиваясь от того факта, что омонимия разрешается фактическим окружением омонима, на этапе её разрешения разработчик семантического парсера автоматически соприкасается с проблемой синтаксиса. В русском языке синтаксические связи предопределяются семантикой повествования, которая в алгоритмической среде может быть смоделирована через семантические категории лексем. Таким образом в вычислительной системе семантические категории слов могут быть назначены экспертом на этапе построения морфо-семантического словаря. И важным моментом здесь является обработка ситуации по назначению семантических признаков конкретной словоформе в морфологической словарной статье. Тогда предназначение программы семантического разбора текста будет состоять в

удалении из набора семантических категорий слова-омонима его избыточных значений. При таком подходе появляется необходимость в назначении лексеме признака, несущего в себе техническую функцию управления алгоритмом разрешения омонимии, а именно признака промежуточного статуса принятия синтаксического решения – «не надо»/«надо»/«было». «Не надо» – семантическая категория в слове единственная, значит разрешение омонимии не требуется; «надо» – семантических категорий в словоформе несколько, значит требуется применение алгоритма разрешения омонимии и, наконец, «было» – семантическая категория ранее была выбрана и обращение к алгоритму разрешения омонимии не требуется вне зависимости от количества оставшихся семантических категорий в словоформе.

Таким образом вычислительный семантический анализ текста сводится к условной расстановке семантических категорий (через отсечение избыточных) над лингвистическими единицами, то есть над словами, устойчивыми оборотами, синтаксическими единицами (именные конструкты, предложные группы, определительные обороты). Семантические категории берутся из морфо-семантического словаря и по необходимости уточняются в соответствии с признаком статуса принятия синтаксического решения. После назначения лексеме синтаксического статуса её семантическое значение можно считать автоматически определённым, если соотнести лексему со структурой семантической сети, основанной на синтаксической иерархии. Семантическая сеть, построенная по описанной методологии, является полноценной базой для проектирования прикладных смыслоориентированных решений.

«Так не говорят!» Аутентичность материала при обучении иностранному языку

Воронцова Марина Игоревна
ЛАЛС НИВЦ МГУ имени М.В. Ломоносова
(Москва)

В докладе обсуждается вопрос использования аутентичных материалов при изучении иностранного языка. Особое внимание уделяется использованию цифровых технологий для освоения лексики.

Нужны ли аутентичные материалы в классе?

Какого рода данные необходимы для уверенного владения лексикой?

Могут ли современные технологии способствовать адекватному использованию слов и выражений?

Вопрос об использовании аутентичных материалов при обучении иностранному языку является предметом неутихающих дискуссий. Высказываются мнения за и против, как например, обилие незнакомых слов на странице, сложность восприятия на слух речи носителей языка или иноязычных говорящих. Тем не менее, при обучении продуктивным навыкам, таким как говорение и письмо, мы неизбежно оказываемся перед необходимостью использовать ту или иную речевую конструкцию, слово или словосочетание в контексте, в реальной жизненной ситуации.

Чтобы повысить уверенность обучающегося в использовании иностранного языка, необходимо познакомить его с тем, как это делают носители языка. Если необходимо написать письмо или заявление о приеме на работу, хорошо бы иметь образцы таких писем и заявлений. Если же надо позвонить по телефону, ответить на звонок, познакомиться, поблагодарить за обед или попрощаться, то помогут «разговорники» или записанные диалоги и монологи, произнесенные по разным поводам и разными участниками (разными по полу и возрасту, социальному положению, происхождению, образованию и т.п.)

Надежный контекст необходим не только для сложных текстов, но и для освоения лексики. Что же делать, чтобы не попасть впросак и не использовать, например, идиому, которая давно вышла из употребления, и не услышать от собеседника: «Так не говорят!»?

- Использовать примеры употребления слов и словосочетаний, которые приводят словари, причем онлайн версии крупных издательств расширяют списки примеров, базируясь на корпусных исследованиях (например, https://www.ldoceonline.com/)
- Учитывать частотность употребления слова или словосочетания, идиоматического выражения, и в зависимости от этого решать, включать или не включать эти языковые единицы в свой лексикон или в программу освоения учащимися данного уровня.

Так например, известная всем идиома «rain cats and dogs» согласно корпусным исследованиям имеет частоту 65, а «get rid of something» - 28000, хотя ее не включали в список обязательной лексики на ранних этапах изучения языка. (https://corpus.byu.edu/bnc/, https://corpus.byu.edu/iweb/)

- Обеспечить учащихся аутентичными материалами разных типов: текстами, аудио и видеозаписями, такими как , радио и телепрограммы, лекции и другие материалами, которые не были специально созданы для учебных целей, а создавались носителями языка для носителей языка или людей, свободно им владеющих. (Например, TED Talks https://www.ted.com/talks)
- Использовать цифровые технологии для поиска необходимого материала, например, нахождения нужной фразы, использованной в естественном контексте, в том числе аудио и видео, как например, https://www.playphrase.me

Таким образом можно подготовить учащихся к уверенному общению на иностранном языке в реальном мире.

ЛИТЕРАТУРА

Gavin Dudeney, Nicky Hockly. How to Teach English with Technology. Pearson Longman, 2007

Lewis Lansford. Authentic materials in the classroom: the advantages. http://www.cambridge.org/elt/blog/2014/05/16/authentic-materials-classroom-advantages/

Michael Lewis. The Lexical Approach: The State of ELT and a Way Forward, 1993

Michael McCarthy. Accessing and interpreting corpus information in the teacher education context.

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.875.716&rep=rep1&type=pdf

Michael McCarthy. No more ‘raining cats and dogs’? An informed approach to teaching English idioms. http://www.cambridge.org/elt/blog/2018/02/16/no-more-raining-cats-and-dogs-an-informed-approach-to-teaching-english-idioms/

Кетский корпус: 1937 — 2018

Ю.Е. Галямина
НИВЦ МГУ им. М.В.Ломоносова
(Москва)

Современные подходы к работе с языками, находящимися под угрозой исчезновения, во многом связаны с методом документирования результатов речевой деятельности (текстов), что неизбежно ставит вопрос о создании корпусов на этих языках. Кетский язык не стал исключением. По нему ведется долгосрочная работа, связанная с записью, расшифровкой и глоссированием и другими типами лингвистического анализа текста.

В результате работы, которую Лаборатория автоматизированных лексикографических систем Научно-исследовательского вычислительного центра МГУ ведет на протяжении без малого 15 лет, собран корпус устных текстов, который превышает 100 единиц. Кроме того, ЛАЛС ведет работу с архивами, собирая тексты, которые не вошли в научный оборот (прежде всего речь идет о тетради Г. Корсакова, в которой собраны устные тексты конца 30-х годов 20 века). Третье направление документирования — создание корпуса письменных текстов, созданных начиная с 90-х годов носителями языка, а также Г. Вернером, в основном в учебных целях.

Остановимся подробнее на корпусе устных текстов, собранных в полевых экспедициях. Очевидно, что они не представляют собой единый массив, а имеют разнообразные характеристики.

1. Диалект. В корпусе предаставлены тексты на южном и северном диалектах, частично встречаются и среднекетские тексты. Основной массив текстов относится к южному диалекту, так как именно на этом языке говорят основные носители языка.

2. Прецедентность/непрецедентрность. По этому признаку тексты делятся на несколько групп. Есть определенное количество записанных прецедентных, традиционных фольклорных текстов (сказок, легенд, песен), которые повторяют нарративные характеристики, связанные с героями, сюжетами и устойчивыми фольклорными формулами. Однако присутствуют в корпусе тексты, которые не носят полностью прецедентный характер, но имеют прецедентные жанровые черты. Например, рассказы об охоте, об аргише или о встрече с медведем. Обычно это личные истории, имеющие собственную сюжетную линию, однако дискурсивная передача событий определяется законами жанра. Другими типами полупрецедентных текстов являются авторские тексты, которые являются прецедентными только для автора: однажды сочиненные они повторяются с определенными вариациями. Третьим типом текстов являются окказициональные тексты, в которых рассказчик рассказывает о своем уникальном опыте, не укладывающемся в жанровые рамки.
3. Третий набор характеристик связан с жанрами. Кетский текстовый мир располагает широким набором жанров, среди которых в нашем корпусе собраны тексты крупной формы (нарративы разных типов: сказки, легенды, истории из жизни, рассказы о шаманах, охотничьи рассказы, рассказы об аргише и т. д.) и более мелкие жанры — шаманские песнопения, современные варианты личных песен.
4. Тексты, собранные в корпусе также характеризуются поколенческими отличиями в языке и речи. Есть тексты старшего поколения (1910-1940-е года рождения), среднего 1950-1960-е года рождения и самого молодого (конец 60-х-70-е года рождения). Язык и языковая культура этих поколений отличаются (вплоть до исчезновения некоторых падежей).
5. Классифицировать тексты можно также по монолгичности/lиалогичности . В основном в корпусе собраны монологи, однако диалогические тексты более всего подходят носителям с неуверенным знанием языка.
6. Еще одним показателем является наличие и уровень смешения кодов (с русским языком), которые позволяют классифицировать тексты от полностью кетских до русских с отдельными кетскими вкраплениями.

Таким образом собранный корпус дает лингвистам и ученым других специальностей, в том числе будущих поколений, наиболее полную информацию о языке, текстах жанрах и динамике изменений кетского языка и речи в начале XXI века. А с учетом архивных материалов и более длительного периода.

В. М. Андрющенко, первый навык филолога и компьютерные филологические системы

С. И. Гиндин
РГГУ
(Москва)

В нашем представлении о развитии отечественной гуманитарии о последней трети XX века Владислав Митрофанович Андрющенко прочно прописан по ведомству вычислительной (теперь говорят:: компьютерной) и прикладной лингвистики.

Между тем его концепция Машинного фонда русского языка подспудно, но тесно связана с идеями и задачами филологии. Данный тезис может показаться тривиальным, если понимать под филологией, как это часто бывает, механическое объединение языкознания и литературоведения. Но если понимать под филологией особую область деятельности и знания со своими специфическими задачами (см.,к примеру, /1/ и названные там работы), от которой науки о языке и литературе давно отделились, то утверждение о связи проекта по прикладной лингвистике с филологией нуждается в разъяснении и обосновании.

В подтверждение неслучайности данной связи для В. М. Андрющенко в докладе предполагается рассказать о двух эпизодах из личного общения с ним. Первый из них касается судьбы рукописей С. И. Бернштейна и позволяет дополнить и уточнить её описание, данное В. В. Золотухиным /2/. Второй связан с помощью Владислава Митрофановича на заключительном этапе проекта по созданию гипертекстовой филологической системы по творчеству В. Я. Брюсова (bryusov.rggu.ru см.также /3/).. Именно при осмыслении итогов.этого проекта возникло понятие систем филологического обеспечения как особого важного класса компьютерных текстовых корпусов

и обобщенной модели результаов труда филологов (см. /4-/ и до.публикации)..

Представляется, что выводы предлагаемого рассмотрения показательны не только для характеристики чёрточек лично В. М. Андрющенко, но и для понимания принципиального родства задач прикладной лингвистики и филологии.

ЛИТЕРАТУРА

1. Гиндин С. И. Что же такое филология и какова ее современная структура // Проблемы поэтики и стиховедения. Алмааты, 2012.

2. Золотухин В. В. Архив С. И. Бернштейна в Доме-музее М. Цветаевой// Бриковский сб. Вып. 2. М., 2014.

3. Гиндин С И. Гипертекстовая филологическая система по творчеству В. Я. Брюсова: Предварительное сообщения// Московский лингвистический журнал. 2003.Т.6. N 2.

4. Гиндин С. И., Иванова Е. А. и др. Системы филологического обеспечения как особая разновидность обогащенных текстовых корпусов // Вестник РГГУ. 2008. N6. Моск.. лингв. журнал. Т.10.

e-mail: sigindin@gmail.com

Корпусная лингвистика в Институте лингвистических исследований РАН: история и современное состояние

Е.В. Головко

Институт лингвистических исследований РАН

evggolovko@yandex.ru

В докладе будет представлена история развития корпусной лингвистики в ИЛИ РАН. Лаборатория автоматизации лингвистических исследований была создана в начале 1990-х годов и, таким образом, разрозненные попытки отдельных исследователей использовать

компьютерные методы для обработки языковых данных получили институциализованное воплощение. Создание Лаборатории вызвало разную реакцию – от полного восторга до глубокого скепсиса. После нескольких лет интенсивной работы наступил спад, на смену которому уже в 2000-х годах, на новом витке развития компьютерных технологий пришла систематическая работа по созданию лингвистических корпусов.

Компьютерные методы в сравнительно-историческом языкознании

А.В.Дыбо, Е.В.Коровина
ИЯз РАН
(Москва)

В докладе будет рассказано о трех основных направлениях, в которых современное сравнительно-историческое языкознание использует компьютерные методы: разработка этимологических баз данных, определение классификации языковых групп и автоматическое сравнение идиомов.

1. Этимологические базы данных.

Сами словарные базы компаративистской направленности стали появляться давно и задолго до компьютерной эпохи. Одна из наиболее известных и до сих пор используемых компьютерных баз - Dyen Isidore, Comparative Indo-European languages (https://thevore.com/ie/cmp/). Сейчас трудно установить время ее появления в компьютерном виде, однако одна из ранних распечаток этой базы данных датирована 1972 г. Это база состоит из 100- и 200-словных списков Сводеша и используется в основном для глоттохронологических расчетов. С 1960-х гг. действует проект Pollex (этимологический словарь полинезийских языков), первым его разработчиком стал Б. Биггс, затем в работе участвовали и другие исследователи (https://pollex.shh.mpg.de/, в настоящее время включает более 5000 этимологий). К настоящему времени словарные этимологические базы данных довольно широко. Однако, в основном, они представляют

собой просто электронные таблицы с небольшими возможностями сортировки и поиска.

Выделяется на этом фоне СУБД STARLing, разрабатывавшаяся С.А.Старостиным и его соратниками с 1986 года (первоначально на основе dBase и Clipper), http://starling.rinet.ru/. Starling создавался, прежде всего, как "рабочее место лингвиста-компаративиста" (см. Старостин 1993), но обладает и массой функций, делающих ее полезной, например, для корпусных исследований (Крылов 2008).

Среди функций STARLing: возможности создания разнообразных выборок; классификация и автоматическое сравнение (о чем см. ниже); автоматическая конвертация текста в базу данных с последующими возможностями корпусной обработки; автоматический морфологический и синтаксический анализ, построение синтаксических деревьев (в частности, впервые, еще в версии 1990 года, был осуществлен морфологический анализ для русского языка на основе словаря Зализняка); возможности стиховедческого анализа. Пожалуй, более многофункциональной системы для обработки языковых данных в определенном смысле до сих пор нет. В настоящее время, так сказать, по мотивам STARLing, осуществляется разработка системы Lingvodoc, которая сможет на новой платформе использовать все функции Starling, плюс некоторые возможности фонетического анализа - см. доклад Ю.В.Норманской на настоящей конференции.

Некоторые специальные возможности поиска и сортировки, необходимые для сравнительно-исторического исследования, предусмотрены и в базе данных RefLex по языкам Африки http://sumale.vjf.cnrs.fr/Lexiques/reflex/RefLex_description.pdf. Это, в частности, возможности поиска по слоговой структуре, консонантному скелету и тональной схеме. Надо отметить, что в СУБД STARLing эти возможности осуществлены программно, в то время, как в RefLex данные для этого поиска заложены как заполняемые вручную поля базы данных.

Разумеется, даже простейшие возможности поиска и сортировки, предоставляемые базами данных, не говоря о возможностях системы STARLing, сильно ускоряют работу компаративиста, позволяя осуществлять одновременно обработку большого числа словарей, что важно как для диалектных исследований, так и для исследований по дальнему родству.

2. Классификации

Впервые формализованный метод генеалогической классификации языков был предложен в 1950 г. М. Сводешом. Основная цель его исходно состояла в получении абсолютной датировки распада языков (узлов генеалогического древа). Однако лингвисты, прежде всего занимающиеся малоизученными языками, сразу стали использовать его просто для получения генеалогического древа, без обращения к абсолютной хронологии. Одной из первых попыток формально разработать методологию построения лексикостатистического языкового дерева была предложена в работе (Sankoff 1969), первая же программа для подобного рода расчетов опубликована в (Guy1980). В докладе будут обсуждаться основные принципы различных методов лексикостатистической классификации и глоттохронологической датировки.

3. Программы автоматического сравнения языков

Программы такого рода появлялись с 60х годов XX в и моделировали в основном задачи установления когнатов, если имеется система соответствий. Позднее они стали пытаться решать и другие классы задач, в частности, нахождение систем соответствий. Мы рассмотрим два основных метода выравнивания когнатов: по консонантным классам, (предложено А.Б. Долгопольским и осуществлено в СУБД STARLing) и система edictor, разрабатываемая М. Листом (http://edictor.digling.org/); существуют и другие разработки, о которых будет кратко упомянуто.

Библиография

Крылов 2008 - *Крылов С.А.* Стратегии применения интегрированной информационной среды STARLing в корпусной лингвистике и в компьютерной лексикографии //Аспекты компаративистики - 3. Orientalia et classica XIX. М., 2008: 649-668.

Старостин 1993 - *Старостин С.А.* Рабочая среда для лингвиста // Базы данных по истории Евразии в средние века, вып. 2. М., Институт востоковедения РАН, 1993: 50–64. / Перепечатано: *С. А. Старостин*. Избранные работы. М.: «Языки славянских культур», 2007: 481-496.

Guy1980 - *Guy J.B.M.* Experimental glottochronology: basic methods and results/ Pacific Linguistics, Series B . No. 75.

Sankoff 1969 - *Sankoff D.* Historical Linguistics as a Stochastic Process. Ph.D. McGill University 1969.

Электронный корпус мансийских текстов: проблемы разработки и перспективы использования

Жорник Дарья Олеговна, Сизов Фёдор Олегович
МГУ имени М. В. Ломоносова, Институт языкознания РАН
(Москва)

В докладе мы представим электронный корпус литературных и диалектных текстов на мансийском языке (http://digital-mansi.com/corpus), а также обсудим некоторые перспективы его использования. Мансийский язык, как и родственный ему хантыйский, демонстрирует высокий уровень диалектной вариативности (см. [Honti 1988]). Хотя на сегодняшний день живыми остаются лишь некоторые северные говоры манси, существует значительное количество текстов и на других диалектах, которые также могут и должны быть включены в корпус. Самые ранние из этих текстов датируются 1840-ми годами, таким образом, мы полагаем, что полный диалектный корпус будет способен отражать изменения в мансийском языке, происходившие на протяжении последних 170 лет. Это позволит использовать его для исследований диахронической эволюции и диалектной дивергенции мансийского языка.

Корпус включает многочисленные коллекции литературных и диалектных мансийских текстов, на базе которых могут быть созданы корпусные словари, распределенные по подкорпусам, каждый из которых соответствует определённой диалектной группе. Тексты XIX и первой половины XX века, представленные в корпусе, были записаны венгерскими, финскими и русскими исследователями (А. Регули, А. Альквистом, В. Н. Чернецовым и др.). Современные литературные тексты представляют собой печатные издания, опубликованные в советский и постсоветский период (они были оцифрованы посредством сканирования и распознавания при помощи программы Tesseract OCR). Другим важным источником литературных мансийских текстов является газета “Лӯима Сэ̄рипос” (http://www.khanty-yasang.ru/luima-seripos), выпускаемая с 1989 года до

настоящего времени. Не менее важен подкорпус текстов, записанных Д. О. Жорник и С. В. Покровской в рамках полевых исследований слабо документированного верхнелозьвинского диалекта [Жорник, Покровская 2017] летом 2017 и зимой 2018 года. Эти тексты включаются в корпус вместе с соответствующими аудиофайлами.

При работе над корпусом мы в значительной степени опираемся на опыт предшественников. Особенно важной для нас является универсальная платформа UniParser, разработанная Т. А. Архангельским [Arkhangelskiy et al. 2012]: в нашем корпусе морфологический анализ производится универсальным парсером AmpEngine, который был разработан Ф. О. Сизовым с использованием некоторых моделей из UniParser. С целью анализа текстов на мансийском языке для AmpEngine был создан “мансийский модуль” на основе грамматики [Ромбандеева 1973]. Он поддерживает обработку текстов с высоким уровнем языковой вариативности — последнее, как известно, является одной из основных проблем морфологической разметки слабо стандартизованных языков (см., например, [Gerstenberger et al. 2017]). Транскрипционные системы, используемые в текстах, могут различаться в зависимости от их диалектной принадлежности. AmpEngine имеет интерфейс для свободного переключения между различными системами записи (как транскрипционными, так и орфографическими).

Представленный корпус может быть, в частности, использован для исследования различных грамматических категорий. К примеру, залог и система дифференцированного маркирования объекта в обско-угорских языках подвержены сильному влиянию со стороны информационной структуры и могут быть эффективно лишь на основе большого объёма текстов (ср. [Kulonen 1989], [Толдова, Сердобольская 2012]). В докладе будут представлены примеры таких корпусных исследований.

Список литературы

Жорник Д. О., Покровская С. В. *Документация верхнелозьвинского диалекта мансийского языка.* Постерный доклад на конференции “Малые языки в большой лингвистике”, МГУ им. Ломоносова, Москва, 2-3 ноября 2017.

Ромбандеева Е. И. *Мансийский (вогульский) язык.* Москва: Наука, 1973.

Толдова С. Ю., Сердобольская Н. В. *Дифференцированное маркирование прямого дополнения в финно-угорских языках* // Финно-угорские языки: фрагменты грамматического описания. Формальный и функциональный подходы. М.: Языки славянских культур, 2012, с. 59-142.

Arkhangelskiy, T.; Belyaev, O.; Vydrin, A. *The creation of large-scaled annotated corpora of minority languages using UniParser and the EANC platform* // Proceedings of COLING 2012: Posters. Mumbai: The COLING 2012 Organizing Committee, 2012. Ch. 9. P. 83–91.

Gerstenberger C., Partanen N., Rießler M., Wilbur J. *Utilizing Language Technology in the Documentation of Endangered Uralic Languages* // The Northern European Journal of Language Technology 4, 2017, pp. 29-47.

Honti, L. *Die ob-ugrischen Sprachen – Die wogulische Sprache.* In The Uralic Languages: Description, History and Foreign Influences, Denis Sinor (ed.), 1988, 147–171.

Kulonen, U-M. *The passive in Ob-Ugrian.* SUS 203, Helsinki, 1989.

Цветы для Владислава Митрофановича

Зайончковская Валерия Петровна
ЛАЛС НИВЦ МГУ
(Москва)

Доклад посвящён воспоминаниям о личности Владислава Митрофановича Андрющенко, основанный на личной переписке за год до его кончины.

Давно известно, что всем юбилярам принято дарить цветы, независимо от того, женщина это или мужчина. Мы дарим этому человеку яркие букеты не только в знак уважения или признания его научных достижений, но и просто так, от широты души, чтобы ещё раз выразить свои добрые чувства. Вот почему в этот знаменательный день все цветы от нашего имени - Владиславу Митрофановичу Андрющенко! Материалом для доклада послужила недавняя моя переписка с Владиславом Митрофановичем в 2017 году.

Russian Corpora B.C.

Захаров В.П.
Санкт-Петербургский университет
(Санкт-Петербург)

Russian Corpora B.C.

Название доклада обыгрывает известную статью «Language Corpora B.C.», написанную Нельсоном Фрэнсисом (Nelson W. Francis), одним из создателей Брауновского корпуса, которая описывает историю формирования и использования словарных картотек в английской лексикографии[1]. Фактически, эти картотеки были прообразами сегодняшних корпусов. Название статьи представляет собой игру слов: Corpora B.C. – корпусы, но не *before Christ*, а *before computers*.

В данном докладе мы хотим сделать экскурс в историю зарождения российского (советского) "корпусостроения" в самом начале корпусно-компьютерной эры. Хорошо известен проект «Машинный фонд русского языка», в рамках которого были созданы первые корпусы русского языка.

Но еще раньше был проект создания частотного словаря русского языка[2], который де факто был сформирован на базе корпуса. Словарь вышел в 1977 г., но основная работа по его созданию была проведена во 2-ой половине 1960-х годов и в начале 1970-х.

Словарь составлен на основании обработки средствами вычислительной техники одного миллиона словоупотреблений, что дало около 40 тысяч единиц словаря. Инициатором и руководителем проекта была Л. Н. Засорина (1929—2016), ею же были разработаны теоретические основы и практическая инструкция обработки лексического материала. В результате была создана

[1] Francis, W. Melson. Language Corpora B.C. In: Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82. Stockholm, 4–8 August 1991, ed. by Jan Svartvik. (Berlin & New York: Mouton de Gruyter, 1991), pp. 17-32.
[2] Засорина Л.Н. (ред.). Частотный словарь русского языка: Около 40 000 слов. М.: Русский язык, 1977.

аналитическая модель переработки сегментов текста в элементы словаря (можно сказать, корпуса), которую можно назвать аналитической грамматикой русского языка (отдельно для каждой части речи) и которая применялась при предмашинной обработке текста.

Машинная обработка велась с помощью счетно-аналитических машин в Вычислительном центре Ленинградского университета. Процесс обработки первой порции текстов в 120000 словоупотреблений был описан в книге Л. Н. Засориной «Автоматизация и статистика в лексикографии». Окончательная обработка материалов была проведена в Горьковском государственном университете на ЭВМ БЭСМ-3М в Лаборатории семиотики НИИ ПМК под руководством В. А. Аграева и В. В. Бородина.

Обобщая опытные данные о применении вычислительных машин в словарной работе, Л. Н. Засорина писала о необходимость изменения организационных принципов словарной работы с использованием ЭВМ и о том, что в этом случае система инвентаризации лексикографических данных превращается в лингвистический банк данных.

В докладе мы хотим познакомить слушателей с документами, подготовленными для обработки лексического материала. Знакомство с методами создания словаря показывает, что уже тогда и там обсуждались все вопросы, которые обсуждаются в корпусной лингвистике. В частности, это проблема репрезентативности и сбалансированности корпуса, или жанровой дифференциации лингвистического ресурса. В качестве источников словаря были определены четыре группы текстов: художественная проза, драматургия, научные и публицистические тексты, газетные и журнальные тексты, при этом драматургические произведения рассматривались как частичный аналог разговорной речи. Также ставилась и решалась проблема определения тождеств и различий речевых единиц, иначе говоря, вопрос об основных единицах корпуса - словоформах (токенах), лексемах (леммах) и словосочетаниях. Особое внимание уделялось задаче разбиения vs. объединения слов, написанных через дефис, который может быть как делимитатором, так и конкатенатором. Также описаны правила обработки точки в текстах корпуса.

Обсуждались и другие проблемы графематического анализа, такие как обработка знаков препинания, эмфатических знаков, кавычек, скобок, подстрочных и надстрочных знаков, чисел, слов, написанных в алфавитах, отличных от кириллического. При этом транслитерированные слова,

заимствованные из других языков, в словник отбирались, а слова бессмысленные (напр., *татути, напотутоте* у С. П. Антонова в «Весне») исключались, равно как и формулы, графики, таблицы.

Создатели словаря отмечали, что наличие машиночитаемой базы словаря (сегодня мы бы сказали, корпуса) позволяет не только создать частотный словарь языка, но и строить обратные словари по отдельным жанрам и источникам, заниматься смысловым анализом лексики, выявлять семантические связи, выбирать метаязыковые формулировки для толкования значений. Машиночитаемый массив словаря рассматривался ими как экспериментальная база для перехода к широкой автоматизации словарных работ.

Таким образом, можно сказать, что в 1960-е годы корпусная лингвистика зарождалась не только в США (создание Брауновского корпуса, 1963-1964 гг.), но и в Советском Союзе, на кафедре математической лингвистики Ленинградского государственного университета.

Квантитативный анализ коннекторов: семантика и функционирование *то есть*[3]

Инькова Ольга Юрьевна
Женевский университет
(Женева)

Надкорпусная база данных коннекторов (НБД), созданная в ИПИ ФИЦ ИУ РАН (подробнее о концепции и возможностях см. [Inkova & Popkova 2017]), позволяет фиксировать и аннотировать реально встретившиеся в текстах употребления коннекторов. НБД функционирует на основе параллельного русско-французского подкорпуса основного корпуса Национального корпуса русского языка (НКРЯ): подкорпус содержит тексты, в основном литературные, общим объемом около 3 млн. словоупотреблений. Разработанная в НБД для исследования коннекторов система аннотирования позволяет получать статистические данные по следующим параметрам их функционирования: выражаемое коннектором отношение, синтаксическая

[3] Исследование выполнено в рамках совместного проекта «Corpus-based contrastive study of connectors in Russian» (РФФИ № 16-24-41002, ШННФ/FNS № IZLRZ1_164059).

природа вводимого им фрагмента текста, позиция коннектора в этом фрагменте текста, порядок следования фрагментов текста, связанных коннектором, статус исследуемой языковой единицы (что важно для полифункциональных единиц, которые могут выполнять в высказывании и другие функции) и (для неоднословных коннекторов) расположение входящих в состав коннектора элементов.

В НБД была сделана сплошная выборка по запросу *то есть* и создано 614 аннотаций, содержащих эту языковую единицу. Программно сгенерированные статистические данные позволяют сформулировать следующие наблюдения:

– *то есть* является коннектором в 592 случаях, т.е. в 96,4% случаев своего употребления;

– в функции коннектора *то есть* устанавливает отношение переформулирования (см. определение в [Инькова, Гурьев 2018]) в 586 случаях из 592, т.е. в 98,9%; остальные случаи приходятся на отношение коррекции;

– порядок следования фрагментов текста, соединяемых *то есть*, всегда p CNT q (где p – первый фрагмент текста, q – второй фрагмент текста, а CNT – коннектор), но *то есть* может занимать во вводимом им фрагменте текста как начальную позицию (95%, 563 случая), так и конечную (3,38%, 20 случаев) и срединную (1,52%, 9 случаев) позицию;

– *то есть* вводит в 48,47% (287 случаев) фрагмент текста без предикации, в 23,6% – с предикацией (140); в 6,6% – сложное предложение в рамках более сложное синтаксической структуры (39); на долю самостоятельных предложений (после точки) приходится 15,5% (92 случая): повествовательное (57), вопросительное (21), восклицательное (14); в 1 случае *то есть* вводит СФЕ; в 5,57% (33 случая) фрагмент текста, вводимый *то есть*, оформлен как вставочная конструкция;

– элементы, составляющие неоднословный коннектор *то есть*, не могут варьироваться (за исключением разговорного варианта *то есь*) и разрываться другими словами, в отличие, например, от элементов, входящих в состав *да и*, которые могут разрываться (*да... и*).

В докладе полученные статистические данные будут проиллюстрированы примерами из НБД; им также будет дана лингвистическая интерпретация. Будет показано, что квантитативный анализ становится неотъемлемой составляющей лингвистического анализа,

дополняя его и являясь средством его верификации, а с другой стороны, обработка статистических данных открывает новые перспективы исследования и дает лингвистам возможность ответить на многие пока не решенные вопросы.

Литература

Инькова О.Ю., Гурьев А.С. (2018), К вопросу о категории пояснения в русской грамматике. *Русский язык в научном освещении*. 2018. № 1. С. 46–73.

Inkova O., Popkova N. (2017), Statistical data as information source for linguistic analysis of Russian connectors. *Informatics and applications*. 2017. Vol. 11, No. 3. Pp. 123–131.

Сочетаемость логико-семантических отношений: количественные методы анализа[4]

Инькова Ольга Юрьевна, Кружков Михаил Григорьевич
ИПИ ФИЦ ИУ РАН
(Москва)

Сочетаемость логико-семантических отношений (ЛСО) – проблема в лингвистике новая. Ей посвящено мало работ и, как правило, приводимые в них гипотезы не подкрепляются количественным анализом. В основном, изучаются данные английского языка на примере сочетаемости сочинительных союзов и единиц, попадающих в русской грамматике в класс «конкретизаторов», причем, в первую очередь, сочетаний показателей в рамках одного ЛСО. Наиболее изучены сочетания показателей противопоставления: англ. *but*, *in contrast, however, rather, instead, nevertheless, despite that* и под., а анализ сочетаний показателей разных ЛСО оставляется «for another day» [Fraser 2013: 320]. Изучаются также отношения между клаузами сложных предложении, имеющих иерархическую структуру, т.н. деревья зависимостей (см., в частности, [Webber 2006, 2016] и базу данных Penn Discourse TreeBank). Вопросы сочетаемости показателей ЛСО в

[4] Работа выполнена при финансовой поддержке РНФ, грант №16-18-10004, 2016-2018.

работах этого направления специально не изучаются, а предлагаемые объяснения наличия в одном высказывании нескольких показателей ЛСО (например, [Grote & al. 1995], [Oates 2000], [Webber& al. 2001], [Webber 2016]), во-первых, не исчерпывают всех случаев, а во-вторых, не затрагивают вопроса о том, какие именно показатели каких именно ЛСО могут сочетаться между собой.

Разработанная в ИПИ ФИЦ ИУ РАН надкорпусная база данных (НБД) показателей ЛСО позволяет, благодаря используемой системе аннотирования [Зацман и др. 2016], во-первых, увидеть, какие показатели ЛСО сочетаются между собой, а во-вторых, благодаря SQL-запросам, получить статистику, позволяющую определить частотность того или иного сочетания. На 6.04.2018 в НБД зафиксировано 419 сочетаний показателей ЛСО, относящихся к разным семантическим группам. Наиболее частотными являются сочетания показателей соединительных ЛСО и ЛСО экстенсиональной генерализации (*и вообще*) – 45 случаев (10,73%), за ними следуют сочетания показателей ЛСО пропозиционального сопутствования и соединительных (*и при этом*) – 43 случая (10,26%). Наоборот, единичными пока случаями представлены сочетания, например, ЛСО «вопреки ожиданиям» и спецификации (*но например*) или ЛСО альтернативы и экстенсиональной генерализации (*или вообще*). Эти данные будут пополняться по мере наполнения НБД, но уже сейчас позволяют сформулировать некоторые закономерности.

В докладе будет представлена статистика, отражающая возможности сочетаний показателей мереологических ЛСО (генерализации, спецификации, исключения и неединственности), и дана ее лингвистическая интерпретация.

Литература

Зацман И.М., Инькова О.Ю., Кружков М.Г., Попкова Н.А. (2016) Представление кроссязыковых знаний о коннекторах в надкорпусных базах данных, *Информатика и ее применения*. Т.10. Вып. 1 (2016). С. 106-118.

Fraser B. (2013) Combinations of Contrastive Discourse Markers in English, *International Review of Pragmatics* 5 (2013). Pp. 318-340.

Grote B., Lenke N., Stede M. (1995) Mar(k)ing Concessions in English and German. *Proceedings of the 5th European Workshop on Natural Language Generation*, Leiden, May, Leiden University. Pp. 11–32.

Oates S. (2000) Multiple Discourse Marker Occurrence: Creating Hierarchies for Natural Language Generation, *Proceedings of ANLP-NAACL 2000*, *April 29 May 4, 2000*. Seattle, Washington, USA. Pp. 41–45.

Webber B., Knott A., Joshi A.K. (2001) Multiple Discourse Connectives in a Lexicalized Grammar for Discourse, *in* Bunt H., Muskens R., Thijsse E. (eds). *Computing Meaning. Studies in Linguistics and Philosophy*, vol 77. Springer, Dordrecht. Pp. 229-245.

Webber B. (2006) Accounting for Discourse Relations: Constituency and Dependency, *in* M. Dalrymple (ed.). *Festschrift for Ron Kaplan*, Stanford, CSLI Publications. Pp. 1-22.

Webber B.L. (2016) Concurrent Discourse Relations, *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016"*, Moscow, June 1–4, 2016.

Текстовые корпусы Лаборатории автоматизированных лексикографических систем НИВЦ МГУ: история и современность

Казакевич Ольга Анатольевна
НИВЦ МГУ
(Москва)

Лаборатория автоматизированных лексикографических систем (ЛАЛС)[5] во главе с ее первым заведующим Владиславом Митрофановичем Андрющенко была одним из тех мест, где вынашивались идеи, нашедшие воплощение в мегапроекте «Машинный фонд русского языка», здесь же делались первые шаги к реализации этого мегапроекта. В 1980-е годы велась дискуссия относительно того, какой компонент машинного фонда важнее – словари или корпусы текстов. В результате было принято соломоново решение развивать оба направления – и корпусное, и словарное. Хотя в ЛАЛС конца 1970-х - 1980-х созданием словарей для машинного перевода и

[5] Она же Лаборатория структурной типологии языков и лингвостатистики (1964 – 1968), она же Лаборатория вычислительной лингвистики (1968 – 1988). См. (Казакевич, Членова 2014).

оцифровкой бумажных словарей занимались существенно больше, чем текстами, программное обеспечение разрабатывалось равно интенсивно по двум направлениям: и для построения словарных баз данных (в том числе и из оцифрованных бумажных словарей) (Л.И. Колодяжная), и для обработки текстовых корпусов (их лемматизации и получения по ним частотных словарей и конкордансов, позднее сюда добавился и автоматический морфологический анализ) (Ж.Г. Аношкина). Главным корпусным достижением ЛАЛС тех лет был оцифрованный и отлемматизированный Саратовский корпус устной речи, ставший первым компьютерным корпусом русских устных текстов. В середине 1980-х в ЛАЛС появляется еще один небольшой (10 тыс. словоупотреблений) корпус устной речи, но не русской, а селькупской. Корпус был автоматически отлемматизирован (программы Ж.Г. Аношкиной с последующей ручной правкой) (Казакевич 1990), и по нему были получены частотные словари (Казакевич и др. 2002). Расширение языкового репертуара оказалось тогда вполне естественным шагом в построении машинных фондов. Если на первой конференции «по проблемам создания машинного фонда данных для автоматизированной системы лексикографических исследований» (1983) речь шла, прежде всего, о русском языке, то вторая (1987) и третья (1989) конференции той же тематики привлекли внимание и к другим языкам тогдашнего СССР[6].

После ухода В.М. Андрющенко и ведущих программистов лаборатории Ж.Г. Аношкиной, Л.И. Колодяжной и Е.Н. Морозовой в Институт русского языка в лаборатории возникли новые корпусные проекты, напрямую с машинным фондом русского языка не связанные.

В настоящее время в лаборатории поддерживаются и пополняются следующие текстовые корпусы:

- Селькупский корпус, единственный из сегодняшних корпусов лаборатории, запущенный еще на большой ЭВМ и перенесенный затем на персональный компьютер. Корпус существенно пополнился (ок. 100 тыс. словоупотреблениий) и стал мультимедийным. Часть текстов снабжена морфологическими глоссами и образует аннотированный подкорпус. Это по-прежнему корпус устной речи, однако начато формирование селькупского корпуса письменных текстов: http://siberian-lang.srcc.msu.ru (О.А. Казакевич).

[6] См., например, сборники (Машинный фонд… 1988; Машинные фонды… 1987; 1988; Вторая всесоюзная конференция… 1988; Материалы III Всесоюзной конференции… 1990) и др.

- Эвенкийский (более 100 тыс. словоупотреблений) и кетский (ок. 50 тыс. словоупотреблений) мультимедийные корпусы устной речи с аннотированными подкорпусами. В стадии формирования эвенкийский и кетский корпусы письменных текстов: http://siberian-lang.srcc.msu.ru (О.А. Казакевич, Е.Л. Клячко, Ю.Е. Галямина).

- Корпус «Поэзия Московского университета»: http://www.poesis.ru (А.В. Рафаева, М.И. Воронцова, И.В. Шумарина, С.Ф. Членова, Э.К. Лавошникова, Н.Т. Тарумова)

- Корпус дневниковых текстов: http://uni-persona.srcc.msu.ru (М.Ю. Михеев).

- Корпус русских волшебных сказок (более 1,3 млн. словоупотреблений, А.В. Рафаева)[7].

- Французско-русские и русско-французские параллельные корпусы (текст – перевод) (Е.Э. Разлогова).

Большинство корпусов лаборатории – это корпусы первого уровня.

В докладе предполагается рассмотреть формирование и использование корпусов ЛАЛС, причем на корпусах трех малых сибирских языков мы остановиться более подробно.

Литература

Вторая всесоюзная конференция по созданию Машинного фонда русского языка. Материалы конференции. М, 1988.

Казакевич О.А. Использование ЭВМ для исследования бесписьменных и младописьменных языков. М.: МГУ, 1990.

Казакевич О.А., Кузнецова А.И., Хелимский Е.А. Очерки по селькупскому языку. Тазовский диалект. Том 3. М.: МГУ, 2002.

Материалы III Всесоюзной конференции по созданию Машинного фонда русского языка М.: МГУ, 1990.

Машинные фонды языков народов СССР. Материалы рабочего совещания (Тбилиси, 15-22 ноября 1987). Тбилиси, 1987.

Машинные фонды языков народов СССР. Материалы рабочего совещания (Таллин, 19-22 декабря 1988). Таллин, 1988.

Машинный фонд русского языка: предпроектные исследования. М., 1988.

[7] Подробное описание дано в (Рафаева 2014).

Казакевич О.А., Членова С.Ф. Полвека лаборатории автоматизированных лексикографических систем НИВЦ МГУ им. М.В. Ломоносова // Вестник РГГУ Московский лингвистический журнал). М.: РГГУ, 2014. Т. 16, № 8 (130). С. 28 -39.

Рафаева А.В. Компьютер-Слово-Фольклор. М.: РГГУ, 2014.

Оцифровка и обработка печатных материалов на малоресурсном языке: проблемы и решения

Клячко Елена Леонидовна
АО «Синимекс»
(Москва)

В работе рассматриваются проблемы, возникшие при оцифровке и постобработке печатных материалов на эвенкийском языке. Рассматриваются программные средства для решения этих проблем. Кроме того, указываются варианты использования обработанных материалов.

Эвенкийский язык, как и другие малые языки, относится к малоресурсным языкам (low-resource languages). Основной источник при формировании корпусов эвенкийского языка — это устные экспедиционные тексты ([2]). Другой источник — это немногочисленные тексты, опубликованные в интернете, в частности в социальных сетях ([4]). Наконец, в проекте «Корпуса ИАЭ РАН» ([3]) привлекаются тексты из литературных произведений и СМИ.

Барьером к полноценному использованию литературных текстов является малый объем оцифрованных данных. При этом, в отличие от многих языков Сибири, эвенкийский может похвастаться обширной литературой. На эвенкийском языке было издано более 200 книг: проза, поэзия, переводы ([5], см. также электронные каталоги РГБ и РНБ). «Насаждавшийся» литературный язык справедливо критикуется [6], однако авторы текстов — это все же носители языка. Кроме того, до сих пор в основном не оцифрованы экспедиционные материалы с текстами, записанными в конце XIX—середине XX века ([7], [8]). И архивные устные, и литературные материалы могут статистически подкрепить или опровергнуть наши

представления об изменениях в эвенкийском языке на протяжении XX—XXI вв.

Проблемы, с которыми мы столкнулись при обработке текстов:

- отсутствие графического стандарта (разные алфавиты, разная орфография при использовании одного и того же алфавита): см. о подобной проблеме в [9].
- отсутствие готовых средств оптического распознавания символов (программный продукт, например ABBYY FineReader [8], часто поддерживает используемый алфавит, но не языковую модель); специфический набор шрифтов
- сложность последующей автоматической обработки (морфологический и тем более синтаксический анализ) из-за большой вариативности текстов

В докладе будут представлены:

- результаты автоматического распознавания текстов с помощью свободно распространяемого пакета tesseract[9], обученного на коллекции эвенкийских текстов
- результаты автоматического морфологического анализа распознанных текстов
- первичные результаты сравнения текстов: современных и архивных, устных и письменных

Библиография

1. Kuzmenko E., Mustakimova E. Automatic Disambiguation in the Corpora of Modern Greek and Yiddish //Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог» (2015). – 2015. – С. 388-398

2. Камаева Е. М. Корпус кетских и эвенкийских текстов // Студенческая сессия Международной конференции «Диалог» (2017) (http://www.dialog-21.ru/media/3990/kamaeva.pdf)

3. Krylova I. et al. Languages of Russia: Using Social Networks to Collect Texts //Russian Summer School in Information Retrieval. – Springer, Cham, 2015. – С. 179-185.

[8] https://www.abbyy.com/ru-ru/finereader/
[9] https://github.com/tesseract-ocr/

4. Функ Д. А., Мамонтова Н. А., Шаховцов К. Г. Электронный корпус фольклорных текстов на языках малочисленных народов сибири (на материалах шорского, телеутского и эвенкийского языков). Принципы создания и структура //Мультимедийные и цифровые технологии в собирании, сохранении и изучении фольклора. – 2011. – С. 162-179.

5. Афанасьева Е. Ф. Эвенки: язык, фольклор, литература, этнография. – 2006.

6. Shirokogoroff S. M. Tungus literary language //Asian Folklore Studies. – 1991. – С. 35-66.

7. Василевич Г. М. Сборник материалов по эвенкийскому (тунгусскому) фольклору //Л.: Изд-во ИНС ЦИК СССР им. ПГ Смидовича. – 1936.

8. Василевич Г. М. Исторический фольклор эвенков: сказания и предания. – Изд-во" Наука", 1966.

9. Кирьянов Д. П., Орехов Б. В., Панова Т. А. Вариативность орфографий в идише и проблема их автоматической транслитерации

Восстановление грамматического эллипсиса при автоматическом синтаксическом анализе русского предложения

Кобзарева Т.Ю. (t.kobzareva@gmail.com)
РГГУ
(Москва)

Язык предлагает множество механизмов, позволяющих при порождении предложения избежать всякого рода повторов, однообразия лексического и структурного: сочинительные конструкции, трансформации для превращения цепочки простых предложений в структуры с придаточными, деепричастными, причастными и другими оборотами, анафорические замещения полнозначных существительных и прилагательных, и в определенных ситуациях лексические повторы позволяет устранить эллипсис.

При автоматическом синтаксическом анализе, служащем фундаментом понимания предложения, результаты работы этих механизмов усложняют задачу. Поэтому восстановление грамматических эллипсисов представляется практически важной задачей.

Грамматический эллипсис, в отличие от семантического, не нарушающего формальной правильности синтаксической структуры, имплицирует синтаксические аномалии: в [Тестелец 2011] автор определяет эллипсис как «невыраженность тех фрагментов предложения, значение которых может быть восстановлено из контекста».

При грамматическом эллипсисе отсутствуют некоторые компоненты текста, которые есть в глубинном представлении и ожидаются в графе поверхностно-синтаксических связей. Т.е. грамматический эллипсис – это отсутствие в предложении каких-то слов или словосочетаний, появление которых мы, соответственно формальной модели синтаксического анализа, ожидаем.

Для построении процедур восстановления эллипсисов, т.е. обнаружения эллипсиса и поиска его антецедента, необходимо решить следующие задачи:

1) исчислить типы грамматических эллипсисов, с которыми мы будем работать;

2) определить особенности контекстов, которые в процессе анализа будут указывать на возможность появления определенного вида эллипсиса;

3) понять, на каком этапе анализа в системе назревает необходимость восстановления эллипсисов определенного типа и когда, соответственно модели анализа, используемой с вистеме, это возможно сделать;

4) построить алгоритм, позволяющий в исчисленных контекстах с эллипсисом находить антецедент.

В нашем случае при поиске эллипсисов и их антецедентов используется только линейная структура контекста. Структура задается на языке **линейных конфигураций,** подобных используемых в [Мельчук 1967].

Будут рассмотрены 4 наиболее распространенных видов грамматических эллипсисов:

Тип I. ***Эллипсис существительных в приглагольных актантах «с сохранением представителя»*** [Падучева 1974]***:*** ситуация, когда для двух или более именных групп элиминирован повтор во второй из этих групп существительного – вершины и еще каких-то слов - с сохранением некоторой компоненты группы, когда для восстановления эллипсиса возможно использовать грамматику поиска антецедента личного местоимения [Кобзарева 2003], работающую при решении задачи установления кореференции имен.

Тип II. ***Эллипсис приглагольных актантов (без сохранения представителя)*** Ситуация, когда опускается целиком именная группа – актант предиката. В русском предложении возможен эллипсис практически любого узла графа предложения или даже нескольких узлов одновременно, часто опускаются первый и третий актанты.

Тип III. ***Эллипсис предикативной вершины в сложносочиненном предложении.***

Ситуация элиминирования при сочинении предложений повтора сказуемого – вершины – во втором (и далее) из сочиненных, когда морфо-синтаксическая структура второго (и далее) в точности повторяет структуру первого из сочиненных.

Тип IV. ***Эллипсис фрагмента предложения с его вершиной и некоторыми актантами***

Ситуация сочинения предложений, структура которых совпадает с точностью до порядка одинаковых лексем в одинаковых формах, где во втором из сочиненных предложений опущены совпадающие компоненты.

Для каждого типа эллипсиса исследуются аномалии и особенности линейной структуры предложения, возникающие при элиминировании потенциального повтора, которые можно использовать при поиске эллипсиса и его антецедента в анализируемых типах эллипсисов. Для каждого типа предлагается алгоритм реализации этих процедур.

Библиография

[Кобзарева 2003] Кобзарева Т.Ю. Проблема кореференции в рамках поверхностно-синтаксического анализа русского языка // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог’2003— М.: Наука, 2003 — С. 278 -284.

[Мельчук 1967] – Мельчук И.А. Автоматический синтаксический анализ. Т. 1. Новосибирск: Ред.-изд. отд. Сибир. отд-ния АН СССР, 1964.

[Падучева 1974] – Падучева Е.В. О **семантике синтаксиса**. Материалы к трансформационной грамматике русского языка. Москва, Изд. Наука, 1974

[Тестелец 2011] – Тестелец Я.Г. Эллипсис в русском языке: теоретический и описательный подходы // Конференция «Типология морфосинтаксических параметров» МГГУ 5.12.2011

О Владиславе Митрофановиче Андрющенко

Колодяжная Людмила Ивановна
(Москва)

Владислав Митрофанович Андрющенко, я думаю, влиял, как руководитель, не только на мою научную жизнь, но и на жизнь, работу, карьеру многих филологов, работавших в созданной им Лаборатории вычислительной лингвистики в МГУ и в отделе Машинного фонда русского языка в Институте русского языка РАН.

Главная идея его научных стремлений – компьютеризация филологических работ, создание с помощью ЭВМ различных компьютерных словарей, получение производных словарей, автоматизация работ с текстами, включая создание частотных словарей, конкордансов и так далее.

В кратких тезисах трудно представить весь грандиозный план его деятельности в этом стремлении.

Фактически он повлиял на создание научного направления в филологии – «Компьютерная лингвистика».

Что касается моей работы под его руководством (с 1973 до 2008 года), я могу только кратко перечислить те проекты, в которых я участвовала не только как программист, но и как филолог.

Одним из первых значимых проектов было создание на ЭВМ двуязычного Англо-русского словаря объемом в 2000 словарных статей. Эта работа проводилась по договору с ВЦП и была успешно сдана в декабре 1980 года.

Именно в эти годы Владислав Митрофанович стал идеологом создания компьютерных словарей и операций над ними с помощью ЭВМ. Именно эта идея легла и в основу моей диссертации, которой он фактически руководил.

В начале 90-х Андрющенко перешел в Институт Русского языка РАН и основал Отдел Машинного фонда русского языка. Замысел, как всегда, был грандиозным. Благодаря этому замыслу, не только проводились ежегодные конференции, но и создавался сам Фонд в виде текстов, словарей, программ их обработки. Я активно участвовала в части создания компьютерных словарей и возможности проведения различных операций над ними.

В этот период была создана компьютерная система UNILEX-D. Благодаря этой системе, были созданы компьютерные версии около 10 русских словарей, включая Словарь синонимов Евгеньевой, Словарь Золотовой, Словарь Орфографический Лопатина, Словарь языка Русской поэзии (под руководством В. П. Григорьева), Словарь синонимов английского языка Вебстера (совместно с Лабораторий Компьютерной лексикографии филфака МГУ под руководством А.А. Поликарпова) и другие автоматизированные версии словарей, работающих в системе UNILEX-D. Проект подробно описан в статье Л. Колодяжной по заданию В. М. Андрющенко в 2005 году. [О Словарях работающих в этой программной системе можно прочитать в статье: cfrl.ruslang.ru›dictionaries.shtm; в Интернет-Википедии в статье Машинный фонд русского языка; статье Л. И. Колодяжная. Научные публикации. http://docplayer.ru/71836145-Kolodyazhnaya-l-i-nauchnye-publikacii.html]

Но… наступила эпоха Интернета.

Тем не менее, работы Машинного фонда присутствуют в Интернете в виде различных текстов русского языка и русских словарей.

Идеи Владислава Митрофановича живут и воплощаются в компьютерной лексикографии до сих пор. В Институте русского языка создается Национальный корпус русского языка, доступ и работа с которым возможна в Интернете.

А мы всегда будем помнить о нашем великом Учителе.

БРУМС как электронный словарь и база данных

Кретов Алексей Александрович

ВГУ

(Воронеж)

Основу словника БРУМСа (Большого русского морфемного словаря) составляет [СССРЛ-1991]. После удаления вариантов словарь насчитывает 165.243 слова.

БРУМС идеологически близок к [RDD-1970] и [СМоРЯ-1986], но полнее их, и морфемное членение в нем применяется не к орфографической форме,

а к этимологической транскрипции – не латинизированной (ср. практику [ЭССЯ]), а кириллической. Традиция нарушена сознательно: отчуждающий эффект латиницы общеизвестен, использование же кириллицы не вызовет отчуждения у широкого читателя, позволив увидеть за пестротой единиц орфографии (графем) строгость и последовательность единиц языка (фонем), за разнообразием форм – единство функции, за тождеством форм – различие функций. Фонемы понимаются по [Бодуэну-1881] в версии его ученика и последователя [Бернштейна-1962].

Глубина этимологического представления лексем задается амплитудой чередования звуков в алломорфах русских слов и соответствует протославянскому (раннему праславянскому) состоянию (примерно 5 в. н.э.) – до начала действия законов восходящей звучности слога и слогового сингармонизма.

Фонемный состав исконной русской лексики, используемый в транскрипции: вокалы А, О, Е, Ы, **Ĕ**; беглые – Ъ, Ь, І; сонанты: Р, Л, Н, М; У/В и И/Й признаются двумя фонемами-сонантами, имеющими слоговой и неслоговой варианты; консонанты: Т, Д, П, Б, К, Г; С, З, Х.

Протезы и эпентезы обозначаются латинскими буквами*: h, j, n, s, v.* Ударные гласные обозначаются заглАвными, префиксальный стык «-», суффиксальный «=», флективный «_», неопределенный «+». Редупликация обозначается точкой: *надежда* {на-дЕ.д=й_а}.

По нашему убеждению, русский литературный язык является восточнославянско-южнославянским симбиозом: *ворота* и *вратарь* в равной мере принадлежат ему. Соответственно, в нём существует две системы правил перехода от глубинной формы {об-дЕ.д=й_а} к поверхностной: восточнославянская *одёжа* [БАС-3, т.13 с.507] и южнославянская *одежда*.

Заимствованные основы не подвергаются морфемному членению (например, *коммунистический* {коммунистИк=ьск_ий}): это дело не языка-реципиента, а языка-источника.

Электронная версия БРУМСа содержит 3 файла: Words, Rools, Links.

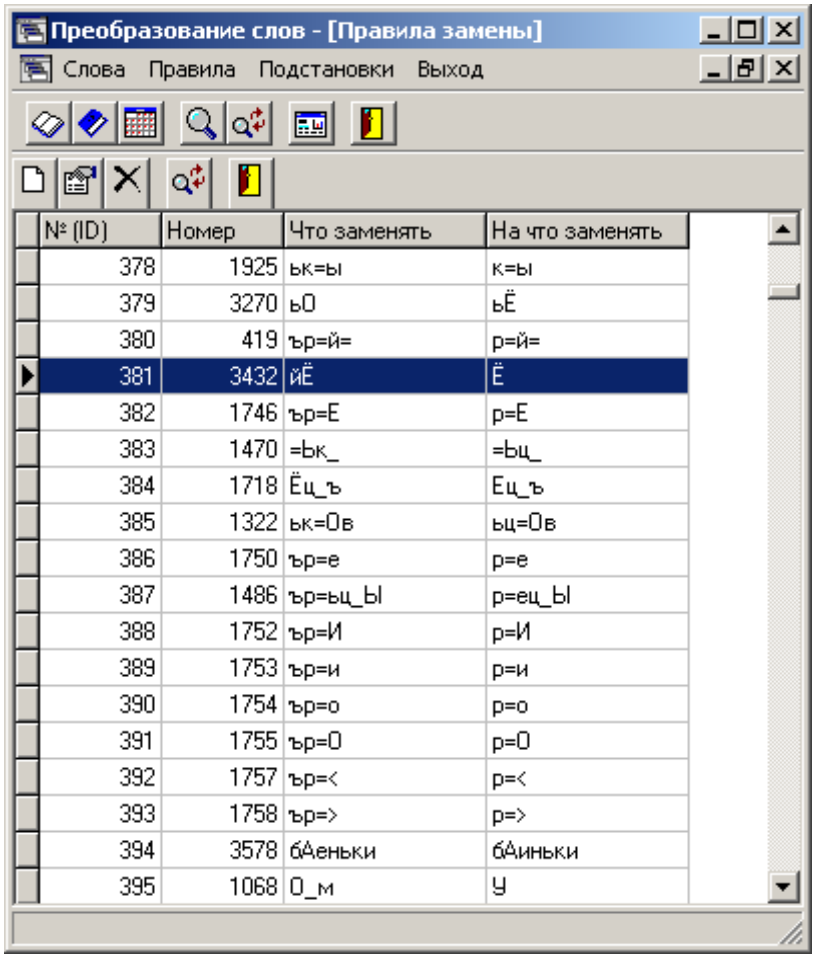
Файл Words (см. Таблицу 2) содержит 4 столбца: ID – имя слова, ЛЕММА – словарная форма слова, СХЕМА – морфемное членение леммы в этимологической транскрипции, ПРАВИЛА ПЕРЕХОДА от Схемы к Лемме – последовательность замен, обеспечивающая переход от этимологической

транскрипции к словарной форме (правила имеют номер - ID, которым они и представлены в четвертом столбце).

Таблица 2. Структура файла Words.

D	ЛЕММА	СХЕМА	Правила перехода от Схемы к Лемме
9353	ОДАРЁННОСТЬ	об-да=р=и=Е=н=ьн=ост_ь	3081;141;723;68;1884;3651;1353;4356;53;54;52;149;130
9354	ОДАРЁННЫЙ	об-да=р=и=Е=н=ьн_ый	3081;141;723;68;1884;3651;1353;53;54;52;1305;2507;
9355	ОДАРИВАНИЕ	об-дА=р=и=ыва=н=ій_е	207;3876;2940;72;650;3651;1353;53;54;52;3857;3866;

Файл Rools (см. Рис.1) также содержит 4 столбца: ID леммы, Номер (ID) правила (порядок применения правил значим), Символы на ВХОДЕ, Символы на ВЫХОДЕ.



№ (ID)	Номер	Что заменять	На что заменять
378	1925	ьк=ы	к=ы
379	3270	ьО	ьЁ
380	419	ър=й=	р=й=
381	3432	йЁ	Ё
382	1746	ър=Е	р=Е
383	1470	=ьк_	=ьц_
384	1718	Ёц_ъ	Ец_ъ
385	1322	ьк=Ов	ьц=Ов
386	1750	ър=е	р=е
387	1486	ър=ьц_Ы	р=ец_Ы
388	1752	ър=И	р=И
389	1753	ър=и	р=и
390	1754	ър=о	р=о
391	1755	ър=О	р=О
392	1757	ър=<	р=<
393	1758	ър=>	р=>
394	3578	бАеньки	бАиньки
395	1068	О_м	У

Рис. 1. Структура файла Rools.

Файл Links (см. Таблицу 2) имеет 3 столбца: ID леммы, ID правила и Prio – реальный порядок применения правила в данном случае.

Таблица 2. Содержание файла Links.

IDWord	IDRule	Prio
110643	48	3096
110840	48	3097
110841	48	3096
111143	48	3100

Правило 48 чаще применяется 3096-ым, но может применяться и 3097-ым или 3100-ым, что придаёт гибкость линейной последовательности применения правил.

БРУМС даёт максимально полную информацию о номенклатуре, дистрибуции, продуктивности морфем, составляющих основы русских слов, необходимую в историко-лингвистических исследованиях, в славянской этимологии (книга [Кретов 2009] – всего лишь заметки на полях БРУМСа), для оптимизации русской орфографии, получения информации об относительной хронологии фонетических и деривационных процессов в русском языке, при отборе материала для обучения русскому языку, а также – при компьютерной обработке текстов на русском языке.

База данных БРУМСа содержит информацию необходимую для синтеза и интерпретации отсутствующих в ней слов.

СПИСОК СОКРАЩЕНИЙ:

RDD-1970 Russian Derivation Dictionary by D. Worth, A. Kozak, D.B. Johnson. – New York, 1970, XXIV+747 pp.

БАС-3 Большой академический словарь русского языка. Том 13 (О-ОПОР). – М.-СПб: Наука, 2009. – 770 с.

Бернштейн-1962 Бернштейн С.И. Основные понятия фонологии / С.И. Бернштейн // Вопросы языкознания, 1962, № 5, С.62-80.

Бодуэн-1881 Бодуэн де Куртенэ И.А. Некоторые отделы «сравнительной грамматики» славянских языков / И.А. Бодуэн де Куртенэ // Избранные труды по общему языкознанию. Том I. – М.: Изд-во АН СССР, 1963, С.118-126.

Кретов-2009 Кретов А.А. Славянские этимологии / А.А.Кретов // Воронеж: Издательство Воронежского государственного университета, 2009. - 364 с.

СМоРЯ-1986 Кузнецова А.И., Ефремова Т.Ф. Словарь морфем русского языка. – М.: Русский язык, 1986. – 1136 с.

СССРЛ-1991 Сводный словарь современной русской лексики [Текст]: в двух томах / Академия наук [АН] СССР. Институт русского языка им. А.С. Пушкина; ред. Р. П. Рогожникова. - Москва : Русский язык, 1991. Том 1. – А-О. - 800 с.; Том 2. – О-Я. - 739 с.

ЭССЯ Этимологический словарь славянских языков. - М.: Наука, 1974-2016. - Вып. 1-40.

Методы и технологии описания структуры многокомпонентных коннекторов[10]

М.Г. Кружков
ИПИ ФИЦ ИУ РАН
(Москва)

Проблема описания структуры многокомпонентных, иначе говоря, неоднословных, коннекторов до последнего времени не являлась предметом специального обсуждения, структура коннекторов мало изучена, отсутствуют критерии для определения линейных границ коннекторов и их компонентов. В рамках корпусного сопоставительного исследования многокомпонентных коннекторов русского языка осуществляется аннотирование этих коннекторов в корпусе параллельных текстов Национального корпуса русского языка.

Для поддержки этой задачи создана специализированная Надкорпусная база данных (НБД, подробнее об этом инструменте см. Зацман и др. 2016; Kruzhkov 2016). Предлагаемый метод аннотирования неоднословных коннекторов в параллельных текстах предусматривает использование двухуровневой фасетной классификации: аннотированию подлежат с одной стороны конкретные употребления коннекторов в текстах корпуса (контекстное аннотирование), с другой стороны аннотируется внутренняя структура каждого коннектора, независимо от контекста его употребления (структурное аннотирование).

При контекстном аннотировании употреблениям коннекторов присваиваются признаки из следующих кластеров: отношения (какое логико-семантическое отношение маркирует коннектор), структура (описывают

[10] Работа выполнена при финансовой поддержке РФФИ, грант №16-06-00070.

особенности структуры предложения), позиция, порядок следования вводимых коннектором компонентов текста, и т.д.

В рамках структурного аннотирования коннектор описывается внеконтекстно, в результате чего структурное описание однажды введенного в систему коннектора ассоциируется со всеми описаниями употреблений этого коннектора, что позволяет существенно сэкономить время разметчиков и избежать расхождений в описаниях. Структурное описание в настоящее время проводится по двум основаниям: во-первых, для неоднословных коннекторов указываются значимые лексические составные элементы (*не, то, более* и др.; некоторые из них могут самостоятельно функционировать в качестве коннекторов, например, *а* или *и*). Во-вторых, для каждого коннектора указывается к какому структурному типу он принадлежит. При этом выделяются следующие типы коннекторов:

- одноэлементные – состоят из одного элемента (например, *и*, *или*, *но*, *а*, *хотя*, *иначе*, и т.д.),
- многоэлементные – состоят из нескольких элементов (например, *и вообще*, *к тому же*, *а особенно*, *но при всем при том*, и т.д.),
- двухкомпонентные – коннекторы, компоненты которых вводят два текстовых фрагмента (например, *даже если бы||то*, *как только||тут же*, *не только||а прежде всего*, *раз уж||так уж*),
- многокомпонентные – коннекторы, компоненты которых вводят более двух текстовых фрагментов (например, *не только||но даже и||и даже*, *хотя||хотя||однако все же*).

Двухуровневая схема аннотирования коннекторов является открытой и по мере необходимости может пополняться новыми признаками.

Использование двухуровнего подхода к описанию структуры неоднословных коннекторов позволяет избежать, с одной стороны, субъективности в присвоении статуса языковой единицы некоторым сочетаниям, оставляя за бортом другие, имеющие тот же статус (см. постановку вопроса в Инькова 2016), с другой, применения морфологической разметки, ориентирующейся на орфографическое слово, т.е. отделяемое пробелом. Наконец, применяемая в НБД система аннотирования позволяет дать более точное представление о системе связующих средств русского языка, их составе и частотности употребления (см. пример такого описания в Кобозева 2017).

Литература

Зацман И.М., Инькова О.Ю., Кружков М.Г., Попкова Н.А. (2016). Представление кроссязыковых знаний о коннекторах в надкорпусных базах данных, *Информатика и ее применения*. Т.10. Вып. 1 (2016). С. 106-118.

Инькова О.Ю. (2016). «К проблеме описания многокомпонентных коннекторов русского языка: *не только… но и*», *Вопросы языкознания* 2. С. 37-60.

Кобозева И. М. (2017). «Коннекторы контактного предшествования во французском и русском языках по данным параллельного корпуса», *Съпоставително езикознание* XLII, 4, сс. 48-62.

Kruzhkov, M. Supracorpora Databases as Corpus-Based Superstructure for Manual Annotation of Parallel Corpora. *CILC2016. 8th International Conference on Corpus Linguistics. EPiC Series in Language and Linguistics*. Vol. 1. EasyChair, 2016. 236-248.

Информационный портал "Автоматизированное рабочее место русиста" (АРМ-Рус) [11]

Крылов С. А.
ИВ РАН
(Москва,)

1. Компьютерные технологии начали использоваться в лингвистике фактически ещё с начала 1950-х годов, в основном, для решения прикладных задач, как-то: машинный перевод, информационный поиск, реферирование, проверка орфографии. Однако до сих пор отсутствует универсальная программная и информационная среда, которая бы эффективно служила ~~бы~~ как русистам-теоретикам (академическим и вузовским русистам), так и русистам-практикам (прежде всего учителям-словесникам). Проект «АРМ-Рус» направлен на создание интегрированной автоматизированной системы для лингвистических исследований, в первую очередь ориентированных на лексикографию и грамматику как ядро и основу всех разделов лингвистической русистики, то есть заключается в создании портала, в котором собраны воедино многие электронные лингвистические ресурсы, имеющие ценность как в теоретических, так и в прикладных исследованиях русского языка. Таком образом, портал выступает в роли автоматизированного рабочего места лингвиста-русиста. При

[11] Работа выполнена при частичной поддержке грантов РФФИ № 17-04-00594-ОГН «Автоматический словарь РУСЛАН: обновленная концепция, новая лексика», РФФИ № 18-012-00650-ОГН «Семантические категории в грамматическом строе РЯ», РФФИ № 17 29-07049-ОГН «Исследование методами искусственного интеллекта системы когнитивных операций, реализуемых в научных текстах» и РФФИ № 17-29-09158-ОГН «Создание корпуса официально-деловых текстов РЯ (ОДКРЯ)».

проектировании портала учтён опыт, накопленный создателями Машинного фонда русского языка, а также ведущими современными специалистами в области компьютерной лингвистики.

Структура портала.

«АРМ-Рус» включает следующие основные разделы.

1. Грамматический раздел ("Метаграмматики" РЯ и его "Гиперграмматика"). Его построение базируется на полнотекстовых версиях ряда грамматик РЯ, в первую очередь:

академических грамматик РЯ (1960, 1970, 1980, 1989);

энциклопедических словарей по русистике (1979, 1997, 2013);

основных вузовских учебников по курсу "РЯ";

классических грамматик РЯ;

обобщающих грамматик РЯ для иностранных читателей.

2. Лексикографический раздел ("Гиперсловарь РЯ"). Здесь обеспечивается доступ к ряду «оцифрованных» словарей РЯ (словарные статьи включают отсылки к традиционным типам словарей - в первую очередь к толковым, идеографическим (в т.ч. синонимическим и антонимическим), неологизмов, языка художественной литературы, деривационным, сочетаемостным, фразеологическим, ономастическим, лингвострановедческим и энциклопедическим словарям), а также к имеющимся в Интернет-пространстве ресурсам: Русскому Викисловарю, порталам "Грамота.ру", "Яндекс.словари" и др. Частью лингвистического обеспечения этого раздела является:

3. Корпусный (текстовый) раздел: («Педагогический корпус русского языка») Он включает следующие подкорпуса:

3.0. Детская речь (устная и письменная) детей от 2 до 18 лет. Устный подкорпус подразделяется на (а) (мультимедийные) аудио- и видеозаписи; (б) транскрибированные записи. Письменный подкорпус представлен (а) школьными сочинениями и (б) эпистолярным жанром.

3.1. Корпус «обязательного» чтения: (1) учебников для средней школы по всем предметам; (2) классической художественной литературы, входящей в "школьную программу"; (3) вузовских и втузовских учебников по различным дисциплинам; (4) действующих государственных законов (Конституция РФ 1993, УК. УПК, Комментарии к ним, нормы гражданского законодательства, правила уличного движения). К ним примыкает:

3.2. «исторический» корпус российских законов Нового времени (свод законов Российской Империи, постановления ЦК ВКП(б), СНК и Верховного Совета СССР, Советская Конституция 1936 и 1977, постановления ЦК КПСС и Совмина, законы СССР), указов, манифестов и обращений русских императоров, а также докладов генсеков ЦК ВКП(б) и ЦК КПСС, президента СССР и президентов РФ.

3.3. Корпус литературы, рекомендуемой для чтения: (1) общих (универсальных) энциклопедий (а) для детей; (б) для взрослых; (2) специальных энциклопедий.

3.4. Корпус "прецедентных" текстов русской культуры, включающий:

(а) фольклорные тексты разных жанров (сказки, анекдоты, частушки и т.п.).

(б) детскую литературу разных жанров.

(в) крылатые слова и выражения, афоризмы, цитаты и т.п.

3.5. Орфоэпический корпус аудио- и видеозаписей русской устной речи.

4. Библиографический раздел. Он включает как базы данных, отражающие библиографию трудов известных ученых-русистов, так и полнотекстовую электронную коллекцию публикаций в области русистики, снабженную гипертекстовой системой указателей (= "Метасловарь РЯ").

Об одном возможном подходе к построению некоторых компонентов машинного фонда русского языка на современном этапе[12]

Крылов Сергей Александрович
ИВ РАН
(Москва)
Семенова Софья Юльевна
ИНИОН РАН, РГГУ
(Москва)

0. МФРЯ: 35 лет спустя

0.1. Концепция и архитектура МФРЯ, разработанная В.М.Андрющенко (на основе интегрирования мнений научного сообщества), является глубоко продуманной и полностью сохраняет свою актуальность даже спустя 35 лет после конференции 1983 года. Согласно этой концепции, МФРЯ включает (1) ≪информационное обеспечение МФРЯ≫ (1.1) иллюстративно-текстовой фонд (ИТФ); (1.2) генеральный словник (аналог Сводного словаря русских словарей Р.П.Рогожниковой); (1.3) автоматический толковый словарь-тезаурус (АТСТ); (1.4) автоматическую документально-фактографическую информационную систему по русистике; и (2) ≪Инструментальное обеспечение МФРЯ≫ , сосредоточенное в процессоре РЯ, включающем (2.1) формальный словарь РЯ, (2.2) формальную грамматику РЯ, (2.3) лингвистические алгоритмы анализа и синтеза текста, а также (2.4) программное обеспечение.0.2. Многие из задач компонента (1.1) сегодня успешно решаются в русле корпусного подхода к РЯ (именно «многие», хотя и не «все», ибо материалы общего ИТФ проектировались как находящиеся в режиме свободного доступа для всех желающих с возможностью формирования дочерних ИТФ). Большинство же пунктов программы Андрющенко (кроме ИТФ) в рамках корпусного подхода не реализованы.

Рассмотрим вопрос: как можно принципиально подойти к решению задач, очерченных в пп. 1.2, 1.3 и 1.4.

[12] Работа выполнена при частичной поддержке грантов РФФИ № 17-04-00594-ОГН «Автоматический словарь РУСЛАН: обновленная концепция, новая лексика», РФФИ № 18-012-00650-ОГН «Семантические категории в грамматическом строе РЯ», РФФИ № 17 29-07049-ОГН «Исследование методами искусственного интеллекта системы когнитивных операций, реализуемых в научных текстах» и РФФИ № 17-29-09158-ОГН «Создание корпуса официально-деловых текстов РЯ (ОДКРЯ)».

1. Реляционные базы данных как инструмент русской лексикографии

Удобным инструментом записи и переработки лексикографической информации является интегрированная информационная среда StarLing, разработанная в конце 1980-х гг. С.А.Старостиным (впоследствии усовершенствованная) и успешно применяемая при решении широкого круга лексикографических задач, в частности при создании т.н. металингвистических баз данных.

2. Металингвистический подход (МЛП) к русской лексикографии

2.0. Специфика МЛП - в том, что **непосредственным предметом** описания становится не столько РЯ как таковой, сколько его "**описание**" (= "**модель**"), "вторичный" объект - результат сознательной (рефлектирующей) научной интеллектуальной деятельности профессионального лингвиста-русиста.

2.1. Типы метасловарей русского языка

Всякий метасловарь есть надстройка над некоторым метатекстом (источнику) (вторичным текстом, т.е. лингвистическим описанием). Метасловари классифицируются: по типам исходных метатекстов-источников: (словарей; грамматик; монографий; статей) и по типам вокабул (словарных входов), описывающих инвентарь: (А) описываемых единиц языка-объекта (морфем, лексем, слов~~оформ~~, фразеологизмов); (Б) используемых единиц метаязыка: «1.Естественного»: (а) лексических (знаменательных) экспликантов (в т.ч. «концептов»); (б) категориальных таксонов (онтологических категорий; семантических признаков); «2.Терминологического» (полуискусственного); «3.Конвенционального» (искусственного) (помет, идеографических значков).

Гиперсловарь есть метасловарь, надстраиваемый над несколькими метатекстами.

3. Построение АТСТ РЯ

3.0. АТСТ РЯ предлагается строить на основе интеграции данных, уже накопленных в русской лексикографии, путём серии последовательных аппроксимаций.

При разработке формализмов для представления лексикографической информации предлагается ориентироваться на формализм словарного комплекса «РУСЛАН», разработанного на рубеже 1990-х и 2000-х годов в НИВЦ МГУ под руководством Н.Н.Леонтьевой и использовавшегося в реально функционировавших (в те годы) системах обработки текста

ПОЛИТЕКСТ и ДИАЛИНГ. Среди его достоинств - многоцелевое назначение, многоуровневый характер, прозрачный метаязык. Словарь, имеющий такую структуру, способен выполнять также функции «формального словаря» как части процессора РЯ, составляющего «инструментальное обеспечение МФРЯ» (в концепции В.М.Андрющенко).

В такую базу данных естественно и гибко интегрируются данные из существующих словарей РЯ – толковых, синонимических, антонимических, ассоциативных, сочетаемостных, семантических (идеографических), а также из металингвистических баз данных, моделирующих исследования по семантике РЯ (как лексической, так и грамматической, но в первую очередь «интегральной») [13].

Тезаурусное описание свадебных причитаний на основе электронных баз данных

Кукушкина Елена Юрьевна
Проект «Живой стилистический словарь русского языка»
(Москва)

В 2000 году под грифом Института языкознания РАН вышла небольшая книга [Никитина, Кукушкина 2000]. Работу над этим словарем тезаурусного типа авторы вели на основе нескольких электронных баз данных, созданных в рамках идеологии Машинного фонда русского языка. Руководителем проекта и создателем его теоретических основ была С.Е. Никитина. Кукушкину больше увлекала работа с компьютером, которая проходила в тесном общении с разработчиком комплекса программ UNIT Ж.Г. Аношкиной и сотрудником Лаборатории автоматизированных лексикографических систем НИВЦ МГУ Т.Е. Реутт.

Позднее тот же тезаурусный подход был положен в основу проекта «Индивид и социум в народной культуре» (2003–2005), которым руководила С.Е. Никитина. В процессе работы над этим проектом Е.Ю. Кукушкиной было подготовлено несколько тезаурусных словарных статей на материале

[13] См., напр., работы А.М.Пешковского, В.В.Виноградова, А.В.Исаченко, Н.Ю.Шведовой, Ю.С.Маслова, Н.Д.Арутюновой, В.Г.Гака, Г.А.Золотовой, В.П.Недялкова, Т.В.Булыгиной, М.В.Всеволодовой, Ю.Д.Апресяна, В.В.Богданова, Ю.С.Мартемьянова, А.В.Бондарко, И.А.Мельчука, Е.В.Падучевой, О.Н.Селиверстовой, М.Я.Гловинской, С.Н.Цейтлин (и их школ).

севернорусских свадебных причитаний. Эти статьи так и остались в рукописи. В 2009 г. вышла книга [Никитина 2009], состоявшая из «вольного авторского текста» (термин С.Е. Никитиной) и тезаурусных словарных статей, относящихся к семантическому полю «Человек», которые были составлены на материале русских духовных стихов.

В предлагаемом сообщении автор возвращается к тезаурусным словарным статьям БРАТ и СЕСТРА, написанным на материале севернорусских свадебных причитаний в рамках подхода, развиваемого С.Е. Никитиной. Так, для брата невесты как персонажа свадебных причитаний характерны следующие черты. 1. Двойственный социальный статус: более высокий, если брат выступает как заместитель отца, что особенно характерно для роли старшего брата по отношению к невесте-сироте, и более низкий, если брат наряду с сестрами выступает как представитель младшего поколения в семье. 2. Ярко эмоциональное отношение к брату со стороны невесты, связанное с её надеждой на брата как на защитника от жениха-чуженина, а также с жизненными связями в родительской семье. 3. Динамичность и ориентация на внешнее пространство, что отражается в целом ряде тезаурусных функций: обращения и эпитеты, относящиеся к тематическому полю «Птицы», *конь* как основной атрибут брата, обилие разнообразных локусов вне дома, в том числе в неосвоенном пространстве, роль защитника и провожатого как основная функция в свадебном обряде. Образ старшего брата невесты индивидуализирован, в отличие от братьев жениха, которых невеста описывает, используя практически только форму множественного числа – *братья*. Совокупный образ братьев жениха представлен в причитаниях достаточно скупо. Братья жениха противопоставлены брату невесты прежде всего по признаку свой/чужой, что отражается в реализации тезаурусной функции Целое: *семья* vs *чужие люди*; в постоянных эпитетах: *родимые* vs *чужие, нареченные*. Резко противоположны метафорические синонимы, которыми пользуется невеста для характеристики родного брата и братьев жениха: родной брат – это *соколочек златокрылый*, *солнца половиночка*, братья жениха – *воронья стая черная*. Идея подвижности в сочетании с идеей чуждости просматривается в сравнении братьев жениха с волками: *То не волки серые, / А нареченные братевки*. На более поздних этапах обряда образ брата жениха индивидуализируется, лексика, используемая невестой при

описании этого персонажа, становится ближе к характеристикам родного брата.

Сущность образа сестры (*сЕстрицы*) в свадебных причитаниях определяется тем, что все сестры проходят один и тот же путь из родительского дома в семью мужа через переходный обряд – свадьбу. Старшая замужняя сестра входит в круг старших родственниц, наряду с *тетушкой, божатушкой* (крестной матерью) и является образцом, которому следует невеста; при этом в родительской семье замужняя сестра занимает положение гостьи, представителя чужой семьи. Младшая сестра входит в круг ближайших подруг невесты, являясь её помощницей и поддержкой (*хожаночка скорая*, *посылочка верная*).

Работа получала финансовую поддержку от РФФИ (проект 17-04-00421).

Литература

Никитина, Кукушкина 2000: Никитина С.Е., Кукушкина Е.Ю. Дом в свадебных причитаниях и духовных стихах (опыт тезаурусного описания). М.: ИЯз РАН, 2000.

Никитина 2009: Никитина С.Е. Человек и социум в народных конфессиональных текстах (лексикографический аспект): Монография. – М.: ИЯз РАН, 2009.

Эффективность спеллера и проблемные словоформы в системных словарях текстового редактора Word

Лавошникова Элина Константиновна

НИВЦ МГУ им. М.В. Ломоносова
(Москва)

Лаборатория автоматизированных лексикографических систем НИВЦ МГУ в её прежней ипостаси стояла у истоков создания русскоязычных спеллеров. Именно её коллективом были оцифрованы и организованы в базу данных ряд словарей русского языка, в том числе «Грамматический словарь русского языка: Словоизменение» А.А. Зализняка, который лёг в

основу автокорректора «Орфо», а затем и системы Microsoft Word – наиболее популярного текстового редактора (но всё ещё не свободного от некоторых недостатков).

При редактировании и многолетних просмотров разного рода текстов мною было обнаружено довольно много низкочастотных или устаревших слов и словоформ, которые текстовым редактором Word 2016 (модификация 2018 года), как и при работе с предыдущими его версиями, пропускаются без пояснений или указаний на возможную опечатку. Некоторые из таких малоупотребительных слов и словоформ приводятся ниже.

В современных пользовательских текстах слова и словоформы из приведённого ниже списка с большой и даже большей вероятностью могут появиться в результате ошибки или опечатки (***пропуск буквы*** внутри слова, ***замена*** буквы, ***перестановка*** букв) в более употребительных словах.

Примеры слов, отсутствующих в нормативном словаре Букчиной, Сазоновой и Чельцовой [1], но пропускаемых Word'овским спеллером без замечаний (они даются перед скобками, в которых заключены подразумеваемые более употребительные слова и словоформы):

вдержка (*в**ы**держка*),

*ве**тр**еть* – фраза «Я всеми ве**тр**ел...» пропускается Вордом без подчёркиваний (*ве**рт**еть*),

вывил (*вы**я**вил, выв**е**л*),

*выме**ш**ать* и *уме**ш**ать* (при сходстве и соседстве на клавиатуре «ш» и «щ» – вместо *выме**щ**ать* и *уме**щ**ать*),

*выск**а**чу* от *выскакать* (*выск**о**чу*),

высочить и *подсочить* (*выс**к**очить* и *подс**к**очить*),

густить (*г**р**устить*),

*за**вя**лю*(*сь*) (*за**яв**лю*(*сь*)),

заушить (*за**с**ушить*, *за**д**ушить*),

отелить (*от**д**елить*, *от**с**елить*),

поветь 'помещение в крестьянском дворе' (*повес**с**ть, повея**я**ть*),

подирать(*ся*) (*под**б**ирать*(*ся*), *под**т**ирать*(*ся*), *п**р**одирать*(*ся*)),

*подр**о**жать* от *дрожать* (*подр**а**жать*, *под**о**рожать*),

подустить (*под**п**устить*, *под**г**устить*),

покакать – «Он ***покакал*** вперёд...» (*по**с**какать*),

*покл**а**няться* (*покл**о**няться*),

*ползад**а*** – «Змея ползад**а** по саду», синтаксический анализ Word’а в этой фразе ошибок не находит (опечатка в слове *ползал**а*** при соседстве на клавиатуре «л» и «д»),

*помстить**ся*** (*пом**е**стить**ся***),

постыть (*п**р**остыть*, *п**оо**стыть*),

посунуться (*п**р**осунуться*, *по**д**сунуться*),

сбирать(*ся*) – «Как ныне сбирается вещий Олег...» (*с**о**бирать*(*ся*)),

тропить(*ся*) (*т**о**ропить*(*ся*)),

улыба – ‘тот, кто много или часто улыбается’ (*улыб**к**а*).

Некоторые из этих примеров (а также многие другие подобные слова) приведены в нашей недавней статье [2]. Там же для иллюстрации даются ***специально сконструированные фразы***. Согласование со словоформами, которые имеются или порождаются в системных словарях, но совпадают с «популярными» ошибками и опечатками, в них нарушено, однако WinWord не замечает в этих законченных предложениях ***синтаксических*** ошибок. Что касается вышеприведённого перечня, то только для существительного *поветь* вместо *повея**т**ь* и для пары <*ползад**а*** – *ползал**а***> (если второй член не ‘половина зала’, а глагольная форма) в некоторых контекстах синтаксический анализ теоретически мог бы выявить несогласованность.

Выводы из вышеизложенного состоят в следующем. При составлении и пополнении внутренних словарей спеллеров желательно выявлять устаревшие и низкочастотные слова, которые могут ***совпасть с искаженным написанием*** (в результате достаточно вероятных ошибок и опечаток) более употребительных словоформ. Такие «подводные камни» иногда «вылезают на поверхность» в списках рекомендуемых программой-подсказкой вариантов исправления. «Проблемные» в этом плане малоупотребительные словоформы можно либо заблокировать в словарях текстового редактора (в Word’е они будут всего лишь подчёркиваться красной волнистой линией как неопознанные), либо помечать в текстах пользователей каким-нибудь особым образом.

Литература

1. *Букчина Б.З., Сазонова И.К., Чельцова Л.К*. Орфографический словарь русского языка. 4-е изд., испр. М.: АСТ-Пресс книга, 2008. 1296 с.

2. *Лавошникова Э.К.* «Проблемные» слова как причина пропуска ошибок при компьютерной проверке орфографии // Текст. Книга. Книгоиздание. 2017. № 15. С. 113–129. DOI: 10.17223/23062061/15/8

Корпусы классической персидской поэзии

Лахути Лейли Гасемовна
ИВ РАН
(Москва)

Классическая персидская поэзия существует как единое поэтическое пространство по меньшей мере с X века. Корпусы текстов персидских поэтов, а затем и поэтические антологии, начали составляться очень рано. Первая из сохранившихся поэтических антологий была составлена в XIII в. Жанр поэтической антологии продолжал развиваться, и с XVI в. их возникло великое множество. Они добавляли к прежним антологиям новые сведения и могли быть организованы по географическому, и/или хронологическому, а также и по тематическому (собрания стихов поэтов-мистиков) принципу; они могли содержать биографические сведения о поэте (тазкере) или только стихи (*Jung*, *Bayāż*). Параллельно создавались рукописные собрания сочинений поэтов (*Kulliyāt*, *Dīvān*), на полях которых часто давались комментарии (шархи).

С этим была тесно связана лексикография. Начиная с XI в. стали составляться толковые словари поэзии, словари «поэтических терминов» (*истилахат аш-шуʿара*), рифм; толковые словари к отдельным произведениям или авторам. Ранние словари, которые содержали большое количество цитат, больше нигде не сохранившихся, стали источником для корпусов произведений ранних поэтов. Первый современный словарь составил вручную крупный иранский ученый А.-А. Дехода. Работа была начата во время Первой мировой войны, публикация закончена в 1975. Переиздан в 1994г. в 15 томах in folio. Словарь организован по алфавитному принципу, он базируется на обширном корпусе персидской поэзии, включает большое количество иллюстративных поэтических цитат, ссылок на ранние

персидские и индо-персидские словари, приводятся и современные значения литературного персидского языка.

Появление компьютерной техники способствовало созданию электронных корпусов текстов важнейших поэтов и электронных словарей, как CD-ROM, так on-line форматов. По мере совершенствования техники эти электронные издания усложнялись и начинали включать все новые возможности. В докладе пойдет речь об электронных корпусах классической персидской поэзии и словарях, об их возможностях и о проблемах, связанных среди прочего с особенностями арабо-персидской графики.

Составление базы данных для исследования структуры писем на примере адресных и приветственных частей писем в Шахнаме

Лахути София Валерьевна
РГГУ, ИВ РАН
(Москва)

Письма, устные послания и связанные с их отправкой и получением коммуникативные ситуации составляют около 10000 строк, т.е. 10% текста поэмы А. Фирдоуси «Шах-наме», истории Ирана от сотворения мира до завоевания Ирана арабами, написанной на основании письменных и устных источников.

Все эти отрывки могут быть исследованы как с точки зрения структуры писем, структуры коммуникативных ситуаций, так и с точки зрения содержания этих фрагментов и их соотношения с целым набором выделенных в ходе работы параметров, как общих для анализа писем, так и специфичных для посланий из художественного произведения.

В ходе работы был определен набор повторяющихся элементов в структуре писем. Данный доклад посвящен составлению базы данных для анализа лексической специфики адресных частей – «унванов» – и приветствий в письменных посланиях, выполненному на основе полного корпуса посланий.

Рассматривается, какие параметры оказались необходимыми для составления базы данных и отличия в их наборе от полной версии базы, а

также специфика автоматического анализа лексики в тексте с арабской графикой.

Гуманитарные дали (проблема гуманитарных знаний в МГУ)

Леонтьева Нина Николаевна
НИВЦ МГУ
(Москва)

МГУ славится не только своими достижениями в естественной и технической области, но и с самого основания он был средоточием знания филологического. Сам М.В. Ломоносов, имя которого носит главный Университет страны и суперкомпьютер, даёт яркий пример совмещения этих двух сторон человеческой деятельности. Учёные, да и простые пользователи давно ждут, что трудозатратные работы по извлечению знаний из естественных текстов будут переданы компьютерам. Но опыт полувековой работы прикладных лингвистов с текстами показал, что только лингвисты не могут справиться с давно обещанной содержательной обработкой текстов.

Потенциал, заложенный В.М. Андрющенко в основанной им Лаборатории, далеко не исчерпан. Да, оцифрованный словарь А.А. Зализняка, с вариациями и добавлениями, остаётся началом всех начал для любых систем обработки текстов. Он «высвобождает» лексический материал для дальнейших манипуляций и вычислений. Начатый тогда же сбор корпусов текстов получил ещё большее развитие. Исследователь-одиночка А.Я. Шайкевич на пересечении этих тем может «осмыслить» содержание большого корпуса. А вот модели семантического анализа текста с более глубоким проникновением в содержание отдельного текста пока не вышли из стадии экспериментальных поисков.

Стремление к «автоматическому пониманию» текстов – как общечеловеческих, так и ограниченных узкой специальностью – поставило много проблем, относящихся к области теоретической семантики. Трудность в том, что у лингвистов и «специалистов» в предметных областях разные и несовместимые взгляды на то, что такое Семантика и что вкладывается в понятие «Смысл текста». А ведь это и глубоко практическая задача:

построить полную и формализованную Базу Знаний в любой специальности нельзя без анализа содержания профессиональных текстов.

Даже слабое знакомство с техническими и точными науками наводит на мысль об их сходстве с проблемами, перед которыми остановились лингвисты. В основном это методы работы с семантическими сетями и сложными графами: ведь структура текста тоже представляет собой сложный, объёмный и многоярусный мультиграф. При этом очевидно, что комплексный анализ текста, включая сжатие содержания, построение индивидуального знания (т.е. «Смысла» для данного пользователя из данного текста) и даже машинный перевод нужны прежде всего специалистам в технических областях.

В МГУ есть все данные, чтобы приблизить момент взаимодействия гуманитариев и технарей в задаче создания структур индивидуального Знания для любого найденного пользователем адекватного текста. Это первый шаг в составе предложенной автором, но не реализованной в МГУ идеи Большой информационной системы (БИС МГУ). Её эскиз был дан в материалах Первой конференции «Машинный Фонд русского языка».

Казалось бы, где, как не в главном вузе страны, ставить и решать такие проблемы как: А) способы связи лингвистических и экстралингвистических знаний, Б) создание единого метаязыка структур для общих и специальных текстов. Эти, а также смежные с ними проблемы занимают умы ведущих учёных ещё с прошлого века. Но такая работа оказалась нереальной в рамках НИВЦ МГУ – по разным причинам.

В НИВЦе МГУ ведутся наряду с точными исследованиями интересные работы филологического характера. Это поддержка и наполнение сайта «Поэзия МГУ от Ломоносова и до …», создание авторских словарей, полевые работы по сохранению исчезающих языков (в форме речевых баз данных), исследования по поэтике. Это и работы с поисковыми системами, и создание словарных баз данных как инструментов семантического анализа текста; но многое держится лишь на энтузиастах. Надеюсь, что объединенными силами нескольких ныне замкнутых коллективов МГУ можно будет приблизить задачу овладения и знаниями из естественных текстов. Дали станут ближе.

«Владислав Митрофанович Андрющенко - научный руководитель и консультант, главный конструктор Машинного фонда русского языка /МФРЯ/»

Лесников Сергей Владимирович
ИЛИ РАН
(Санкт-Петербург)

В конце 1980-х годов активизировались исследования по русской филологии с применением ЭВМ в рамках проекта по созданию МФРЯ [ТЗ и Координационный план 1985], который формировался как система комплексной автоматизации лингвистических исследований и предусматривал «накопление на машинных носителях всего лексического богатства русского языка, создание фонда лингвистических алгоритмов и программ, фонда полностью завершенных систем автоматического анализа и синтеза русского текста, нескольких информационно-справочных систем по языкознанию» [Андрющенко 1986, 8].

Впервые мысль о МФРЯ высказал академик А.П. Ершов в докладе «К методологии построения диалоговых систем: феномен деловой прозы» 26 сентября 1978 года на научной конференции «Диалог-78»: «Любой прогресс в области построения моделей и алгоритмов останется, однако, академическим упражнением, если не будет решена наиважнейшая задача создания МФРЯ. Это фундаментальная проблема, решение которой будет иметь очень большую научную, общекультурную и прикладную ценность. Не мне, конечно, составлять спецификацию такого фонда, но думается, что по крайней мере он должен содержать полный словарь и генератор словоформ, а также формализованный толковый словарь (тезаурус) русского языка» [Ершов, 115]. Проблемам создания МФРЯ было посвящено три Всесоюзные конференции (1983, 1987, 1989), где речь шла о предпосылках создания, основных проектных решениях и, следуя принципу интеграционного подхода [Андрющенко 1989], о практическом воплощении МФРЯ.

В рамках проекта под руководством В.М.Андрющенко разрабатывались 9 фондов-составляющих МФРЯ: 1) Генеральный словник, 2) Словарный, 3) Текстовой, 4) Грамматический, 5) Терминологический, 6) Диалектологический, 7) Исторический, 8) Фонетический, 9) Лингвистический программно-источниковый [Лесников 2002, 21].

С 1991 года до 1996 выделились направления: 1. Создание и совершенствование сервера ИРЯ РАН и МФРЯ; (в 1995-1998 сервер функционировал, но позднее пришлось отказаться в пользу независимых провайдеров); 2. Полномасштабные испытания систем обработки лингвистических данных UNILEX для Орфографического словаря и Словаря поэзии XX в.; 3. Ускоренное накопление новых источников на основе электронных изданий газет и сканирования произведений русской классической литературы. Полный архив источников МФРЯ к 2005 составлял более 100 млн. словоупотреблений; 4. Участие под руководством Ю.Н. Караулова в работе над словарем языка Ф.М. Достоевского.

После 1992, когда началась поддержка научных исследований РГНФ и РФФИ, - развитие МФРЯ приобрело более упорядоченное очертание: 1. Включение МФРЯ в Интернет (www.irlras-cfrl.rema.ru, www.artint.ru/cfrl, www.tractor.de, cfrl.ru). С 2012 на сайте МФРЯ.РФ. 2. Накопление источников для дистрибутивно-статистического исследования русской прозы последней трети XIX в. и газет конца XX в. 3. Накопление и анализ дистрибутивно-статистических данных, подготовка публикаций сводных данных. 4. Разработка технологии комбинированных изданий продуктов МФРЯ (книга, CD, Интернет). 5. Разработка лексической поисковой системы, способной заменить традиционные словарные картотеки (cfrl.ruslang.ru).

На наш взгляд, представление филологических материалов в компьютерной форме и внедрение современных методов научного анализа, основанных на применении цифровых технологий в русской лексикографии, позволяет взглянуть на эти проблемы по-новому и, в частности, создать общий гипертекстовый свод лексики русского языка.

Литература

Андрющенко В.М. МФРЯ: Идеи и суждения. М.: Наука, 1986.

Андрющенко В.М. Вычислительная лексикография. Её возможности и перспективы // ВЯ.1986. №3. С.42-53.

Андрющенко В.М. МФРЯ: Интеграционный подход. М., 1989. 79с.

Ершов А.П. Методологические предпосылки продуктивного диалога с ЭВМ на естеств. яз. // ВФ. 1981. №8. С.115.

Лесников С.В. Словарь русских словарей: более 3500 источников. М.: Азбуковник, 2002. 334с.

Техническое задание на создание МФ РЯ. М.: ГКНТ и АН СССР, 1985. 31с.

Координационный план по выполнению задания 06.01 “Создать МФ РЯ” научно-технической программы 080.18 на 1986-1990 гг. (Задания и этапы программы на 1985-1990). М.: МИНВУЗ СССР, НТС, 1985. 22с.

Координационный план по выполнению задания 06.01 “Создать МФ РЯ” научно-технической программы 080.18 на 1986-1990 гг. (Задания и этапы программы на 1986-1990) М.: ГКНТ, АН СССР и Минвуз СССР, 1986. 13с.

Решение Координационного Совета и Второй Всесоюзной конференции по созданию МФ РЯ (3 июня 1987). М., 1987. 7с.

Количественная оценка степени сходства неатрибутированного текста с текстами его возможных авторов из числа русских классиков

Михеев Михаил Юрьевич, Эрлих Лев Исаакович
НИВЦ МГУ им. М.В. Ломоносова
(Москва)

В работе рассматриваются критерии близости неатрибуированного текста (текста с сомнительным авторством) – к массивам одного или нескольких достоверно известных авторов. В частности,

1) спорной части романа "Тихий Дон" – к массивам текстов М.Шолохова, Ф.Крюкова или кого-то еще из 20 авторов XIX-XX вв., имеющихся в электронной базе Национального Корпуса русского языка (НК),

2) романа "*Двенадцать стульев*" (**12 ст**) – к текстам Ильфа и Петрова (**ИлП**, но уже без 12 ст) или – Булгакова: здесь проверяется гипотеза, будто сам Булгаков какое-то время был в *литературных неграх* у Ильфа и Петрова (И.Амлински 2013),

3) неоконченного романа "*Они сражались за родину*" (**ОсР**) – к текстам Шолохова (**Ш**, но уже без ОсР) или – Платонова (это версия израильского исследователя З.Бар-Селлы),

4) повести "*Роман с кокаином*" (**РсК**) – к текстам В.Набокова или же практически неизвестного автора М.Агеева/Марка Леви... (авторство РсК более 30 лет назад приписывал Набокову – Никита Струве).

Мы хотим количественно измерить стилистические различия тестов, опираясь на частоты *малых слов*, т.е. вспомогательных, служебных средств языка – **союзов, частиц, дискурсивных и модальных слов, предложных и наречных групп** (но также и некоторых вполне значимых прилагательных, существительных и глаголов).

Тем самым дается критерий количественной оценки близости к неатрибуированному тексту группы достоверных авторов.

Финско-русский и русско-финский параллельный корпус в составе НКРЯ: принципы формирования и перспективы развития

Мищенкова Карина Олеговна (karinam6@mail.ru)
Телеканал “RT en Español” (АНО «ТВ-Новости»)
Москва

Финско-русский и русско-финский корпус становится шестнадцатым параллельным корпусом в составе Национального корпуса русского языка. С момента начала работы над проектом в 2016 году и по мере подготовки к выходу из «инкубатора» в открытый доступ объём корпуса возрастает и приближается к отметке в миллион словоупотреблений. При формировании корпуса соблюдаются принципы репрезентативности и сбалансированности. Корпус представлен текстами научной и художественной литературы (в т. ч. мемуарная проза), публицистическими текстами (статьи оригинальных финских изданий и их переводы, подготовленные проектом ИноСМИ), официальными документами (служебные записки, телеграммы, договоры, конвенции, резолюции, директивы, соглашения, протоколы, постановления, дипломатические ноты). В перспективе планируется формирование в составе рассматриваемого корпуса субкорпусов по регистрам языка.

Языком оригинала текстов, входящих в состав корпуса, преимущественно является финский язык. Среди оригинальных текстов на русском языке с переводами на финский язык были размечены два тома «Ракового корпуса» А. И. Солженицына (145 734 словоупотреблений), а также некоторые мирные договоры, подписанные СССР и Финляндией. Так, исходя из представления, что победители диктуют условия договора, для Тартуского договора 1920 г. оригинальным языком признан финский язык,

для Московского (1940 г.) и Парижского (1947 г.) договоров — русский. В планы работы входит расширение русско-финской части корпуса.

К числу основных достижений работы над корпусом относится установление сотрудничества с коллегами из Школы языка, перевода и литературных штудий Университета Тампере, которые с 1999 года [1] занимаются разработкой финско-русских и русско-финских параллельных корпусов. Среди корпусных проектов группы исследователей Университета Тампере заслуживают внимания корпуса текстов художественной литературы *ParFin* (1 464 584 словоупотреблений) и *ParRus* (1 682 312 словоупотреблений) и корпус государственных договоров *PEST* (552 190 словоупотреблений). Сотрудничество с финскими коллегами заключается во взаимном обмене материалами независимо от объёма размеченных текстов каждой из сторон. Кроме того, на этапе подготовки финско-русского и русско-финского корпуса НКРЯ к выгрузке материалов создатели корпусов *ParFin* и *ParRus* обеспечили возможность тестирования текстов на своей платформе https://puolukka.uta.fi/ [2].

Принципиальным отличием подхода к подбору текстов для пополнения корпуса НКРЯ в сравнении с корпусами Университета Тампере является включение финноязычных текстов ингерманландских и карельских авторов. Романы Юхани Конкка (Урхо Торикка) «Мы-герои. История Карельской авантюры Финляндии (1921-22)» и «Огни Петербурга» были изданы в Хельсинки, однако несут черты ингерманландского варианта финского языка. В качестве важного материала для изучения интерференции финского и карельского языков выступают финноязычные произведения авторов Советской Карелии Антти Тимонена, Ортьё Степанова, Николая Яаккола и др. В корпусе также нашли отражение тексты уроженцев карелоязычных областей востока Финляндии таких как роман Майю Лассила «Пирттипохья и ее обитатели». Героями романа Юхани Ахо «Красная черта» становятся карельское и финское население Финляндии, а также карелоязычные коробейники, переходящие финляндско-российскую границу. Многогранный подход к формированию корпуса НКРЯ играет значительную роль в сохранении текстов на национальных языках России, а также делает корпус эффективным инструментом для изучения языковых контактов в прибалтийско-финском ареале.

Библиография

1. Mikhailov M., Cooper R. Corpus Linguistics for Translation and Contrastive Studies: A guide for research. Routledge, 2016 – 234 p.

2. Finnish-Russian and Russian-Finnish parallel corpora of literary texts https://puolukka.uta.fi/

Машинный фонд русского языка: взгляд В. М. Андрющенко из начала 2000-х

Морозова Елена Николаевна

Институт русского языка им. В. В. Виноградова РАН

(Москва)

Доклад основывается на нескольких текстах В. М. Андрющенко, написанных в 2000–2005 годах и предназначенных для сайта Машинного фонда русского языка (МФРЯ). Ниже приводится цитата из текста «Машинный фонд русского языка в развитии» (своего рода предисловия к сайту), относящаяся к деятельности отдела МФРЯ в постсоветское время:

> После 1991 г., когда нарушилась старая система финансирования и координации научно-исследовательских работ и до 1996 г., деятельность Отдела Машинного фонда русского языка приобрела несколько хаотический характер. <…> Тем не менее можно выделить главные направления деятельности отдела того времени:
>
> 1. Создание и совершенствование сервера Института русского языка и Машинного фонда русского языка в Интернет; <…>
>
> 2. Полномасштабные испытания систем обработки лингвистических данных UNILEX путем участия в подготовке Орфографического словаря и Словаря поэзии XX в.;
>
> 3. Ускоренное накопление новых источников на основе электронных изданий газет и сканирования произведений русской классической литературы. <…>
>
> 4. Участие под руководством Ю. Н. Караулова в работе над словарем языка Ф. М. Достоевского.
>
> <…> когда началась поддержка научных исследований различными фондами <…> развитие Машинного фонда русского языка приобрело более упорядоченное очертание. Выделились <…> направления:
>
> 1. Включение МФ РЯ в Интернет <…>,

2. Накопление источников в целях широкого дистрибутивно-статистического исследования русской прозы последней трети XIX в. и газет конца XX в.,

3. Накопление и анализ дистрибутивно-статистических данных, подготовка публикаций сводных данных <…>,

4. Разработка технологии комбинированных изданий продуктов МФ РЯ (книга+CD+Интернет),

5. Разработка лексической поисковой системы, способной заменить традиционные словарные картотеки.

Более отдаленные перспективы развития Машинного фонда русского языка мы связывали с полным осуществлением концепции, выработанной 1-ой Всесоюзной конференцией по созданию Машинного фонда русского языка 1983 г. В основе этой концепции лежали две главные задачи:

1. Создание компонентов лингвистического обеспечения задач информатики и

2. Информатизация научных исследований в русистике.

Уже тогда было понятно, что эти задачи взаимосвязаны: создание компонентов лингвистического обеспечения задач информатики силами профессиональных лингвистов возможно только при условии информатизации русистики. В то же время информатизация русистики требует использования всех достижений прикладной (вычислительной) лингвистики. Но использование достижений прикладной лингвистики в русистике может быть осуществлено только на достаточно богатой источниковой базе и должно быть нацелено на выдачу результатов в полиграфической форме. В соответствии с этим логика развития Машинного фонда русского языка должна быть выстроена так, чтобы создавались прежде всего базовые компоненты (источники на машинных носителях и в базах данных, лингвистические программно-источниковые пакеты, компьютерные технологии подготовки научных трудов). Эти направления никогда не исчерпают себя, так как источниковая база русистики бесконечна, а программные средства и технологии требуют постоянного совершенствования и обновления. Вместе с тем необходимо расширять поле разработок, сейчас — в направлении реализации методов дистрибутивно-статистического анализа и накопления лингвистических ресурсов в Интернет.

Сейчас, в 2005 г. мы должны признать, что данное научное направление (информатизация русистики) оказалось нежизнеспособным в современных организационно-финансовых условиях и постановка задачи создания Машинного фонда русского языка на ближайшую перспективу должна быть еще более сужена до двух-трех частных задач:

1. Дальнейшее накопление источников на сайтах Фонда и совершенствование его технической базы

2. Дальнейшее развитие функций Автоматической словарной картотеки Фонда

3. Конструирование глобальной лингвостатистической обработки всех текстовых источников Фонда в интерактивном режиме (образец такой обработки представлен в Статистическом словаре языка Достоевского).

На пути решения данных задач появились словари:

А. Я. Шайкевич, В. М. Андрющенко, Н. А. Ребецкая. Статистический словарь языка Достоевского. М., 2003.

А. Я. Шайкевич, В. М. Андрющенко, Н. А. Ребецкая. Статистический словарь языка русской газеты (1990-е годы). М., 2008.

А. Я. Шайкевич, В. М. Андрющенко, Н. А. Ребецкая. Дистрибутивно-статистический анализ языка русской прозы 1850—1870-х гг. Т. 1. М., 2013; Т. 2. М., 2016.

Виртуальная лаборатория ЛингвоДок[14]

Норманская Юлия Викторовна
ИЯз РАН
Москва

В настоящее время создана виртуальная лаборатория Лингводок (lingvodoc.ispras.ru), ее лингвистическая составляющая была разработана под рук. Ю.В.Норманской, а программная отделом О.Б.Борисенко (ИСП РАН). В ЛингвоДоке возможна как одновременная он-лайн работа нескольких участников над одним словарем, так и работа без интернета на домашнем компьютере в десктопной версии с возможностью отложенной двухсторонней синхронизации данных с центральным сервером без конфликтов между версиями данных разных исследователей. Данная программа позволяет использовать сеть Интернет лишь для первичной настройки сразу после установки программы, а в дальнейшем работать полностью автономно. Построение модели данных таково, что полностью исключает возможные конфликты между данными разных исследователей. Есть возможность создания словарей и текстов, у которых есть разработчики разных уровней руководитель и подчиненные, при этом руководитель имеет

[14] Работа выполнена при поддержке гранта РНФ № 15-18-00044 «Информационная система для описания малочисленных языков народов мира. Создание описаний алтайских и уральских языков России, находящихся на грани исчезновения».

возможность полностью контролировать всю работу подчиненных и принимать решения о публикации тех или иных данных. В виртуальной лаборатории ЛингвоДок есть возможность работы с этимологическими диалектными аудиословарями и текстами.

В словарях для каждого слова из словаря можно дать следующую информацию:

1) Словарные записи слов
2) Транскрипции регулярных форм
3) Парадигмы слов
4) Переводы слов
5) Транскрипции парадигм
6) Переводы парадигм
7) Этимологические связи с другими словами
8) Аудиозаписи слов
9) Аудиозаписи парадигм
11) Разметка спектрограмм, выполненная в программе Праат
10) Описания словарей
11) Данные об авторах
12) Данные об организациях, в которых работают авторы

Есть возможность анализа данных представленных по спектрограммам Праата: создание трехмерной системы гласных на основании обсчета формант, анализа данных по долготе и интенсивности относительной и абсолютной (это возможно выполнить при открытии нужного словаря и выборе опции Tools > Phonology). Есть также опция Tools > Statistic, которая позволяет отслеживать с точностью до часа активность каждого из участников он-лайн работы.

Тексты размещаются на сайте в программе Элан и есть возможность автоматической конвертации текстов со звуком и без звука в (аудио)словари, которые в дальнейшем можно сливать с уже созданными словарями из других корпусов текстов на этом диалекте, либо собранных в качестве изолированных произнесений. Ранее подобное программное обеспечение в мировой лингвистике разработано не было. Оно открывает широкие перспективы для взаимодействия специалистов по корпусной лингвистике между собой, поскольку в этом программном обеспечении будет возможна конвертация глоссированных текстов из практически любых форматов Элан в словари, которые возможно в дальнейшем сливать с уже созданными

словарями на этом диалекте или соединять с другими с помощью этимологических связей. В ЛингвоДоке есть возможность вывода всех употреблений той или иной морфемы в любом из словарей или корпусов. Это дает позволяет в будущем объединить весьма большие массивы текстов в единой лаборатории и дать им комплексное осмысление с помощью морфологических и лексических аудиословарей, которые будут объединены этимологическими связями.

Помимо этого в ЛингвоДоке существует возможность сколь угодно сложных запросов поиска и отражения их **на карте** мира http://lingvodoc.ispras.ru/map_search. В настоящее время в системе зарегистрированы и работает около 100 ученых лингвистов из большинства крупных городов России и 6 стран Европы (Германии, Австрии, Финляндии, Швеции, Эстонии, Венгрии).

Корпусная экосистема Школы лингвистики НИУ ВШЭ

Орехов Б. В.

НИУ ВШЭ
(Москва)

Коллектив Школы лингвистики факультета гуманитарных наук Национального исследовательского университета «Высшая школа экономики» начал формироваться в 2011 году. С самого начала в нем состояли специалисты по созданию корпусов, в т. ч. принимающие активное участие в разработке НКРЯ Е. В. Рахилина, А. А. Бонч-Осмоловская, С. Ю. Толдова, О. Н. Ляшевская, Т. А. Архангельский. В 2011 году заметный толчок развитию корпусов различных языков дала Программа фундаментальных исследований Президиума РАН «Корпусная лингвистика», активное участие в осуществлении которой принимали названные лингвисты.

К этому моменту в распоряжении коллектива был корпусный менеджер, разработанный в 2000-х годах для Восточно-армянского национального корпуса: http://eanc.net/ На платформе этого менеджера ещё в рамках Программы Президиума РАН были размещены корпуса бурятского, калмыцкого, татарского, казахского, албанского, лувийского, монгольского, цыганского языков (все размещены на ресурсе http://web-corpora.net/)

[Архангельский 2012]. По окончании Программы все они поддерживаются сотрудниками ШЛ НИУ ВШЭ.

Использование единообразной платформы позволило выстроить последовательную работу по сбору коллекций текстов и по подготовке инструментов разметки для других языков. В ходе учебно-научной деятельности в рамках сотрудничество преподавателей и студентов на домене http://web-corpora.net/ в течение 2012-2017 годов размещены корпуса удмуртского, новогреческого, тайского, амхарского языков, языка идиш, а также башкирский поэтический корпус [Arkhangelskiy 2014]. Подготовленные для этих корпусов текстовые коллекции не только доступны в интерфейсе поиска корпуса, но и служат полигоном для осуществления учебной проектной работы: студентами разработаны инструменты для снятия морфологической неоднозначности для новогреческого и языка идиш, морфологический анализатор для амхарского языка, основанный на технологии машинного обучения, апробированы методы автоматического поиска когнатов в близкородственных языках.

Ключевая работа по поддержке корпусной инфраструктуры проделана Т. А. Архангельским, который, являясь экспертом по работе с корпусным менеджером также является разработчиком платформы универсального языконезависимого морфологического анализатора UniParser [Архангельский 2014], используемого для разметки удмуртской, албанской, казахской и других текстовых коллекций.

Унифицированный характер интерфейса корпусной платформы позволил создать универсальную веб-страницу для одновременного доступа ко всем развернутым на сервере http://web-corpora.net/ корпусам. Индексальная страница сайта позволяет не только перейти к нужному корпусу, но и осуществить быстрый поиск по точной форме в любом интересующем нас корпусе.

В рамках проекта «Языки России» (http://web-corpora.net/minorlangs/) собраны коллекции текстов для будущих корпусов, которые будут обладать богатой социолингвистической разметкой.

Одновременно с созданием корпусов для не охваченных языков идет работа по совершенствованию НКРЯ. Разработаны (пока не внедрены по организационным причинам) современные средства визуализации выдачи, составлены скетчи (http://linghub.ru/RNC_sketches/), идет работа над алгоритмами снятия морфологической неоднозначности в русском языке.

В ноябре 2017 года Т. А. Архангельским представлена первая версия нового корпусного менеджера, «Цакорпус», при разработке которой учтены все недостатки использовавшейся ранее платформы. Развертывание корпусов с помощью нового менеджера должно происходить быстрее и проще. Новые корпуса будут размещаться на новом домене ШЛ НИУ ВШЭ, linghub.ru, который, как следует из названия, будет аккумулировать различные ресурсы компьютерно-лингвистической природы, не ограничиваясь корпусами.

Литература

Т. А. Архангельский. Электронные корпуса албанского, калмыцкого, лезгинского и осетинского языков // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2012. № 4. С. 24-29

T. Arkhangelskiy, M. Medvedeva. An online annotated corpus of Udmurt language // 4th Mikola Conference on Lexicology and Lexicography of the Uralic and Siberian languages. Szeged, Hungary, November 14–15, 2014

Т. Архангельский. Система морфологического анализа текстов UniParser // Конференция по компьютерной и когнитивной лингвистике ``TEL-2014''. Казань, 6–9 февраля 2014

Мультимедийный корпус языка идиш // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2015. № 3. С. 18–25

T. Arkhangelskiy, E. Kuzmenko. Composing the Corpus of Modern Greek: features and methods // 12th International Conference on Greek Linguistics. Berlin, Germany, September 16–19, 2015

Т. Архангельский, Т. Панова. Мультимедийный корпус языка идиш // XX ежегодная международная конференция по иудаике "Сэфер". Москва, 12 июня 2015

T. Arkhangelskiy. Digital corpora at web-corpora.net: Features and development // 8th Tsakonian conference. Leonidio, Greece, September 9–11, 2016

T. Arkhangelskiy, M. Medvedeva. Developing Morphologically Annotated Corpora for Minority Languages of Russia // CLiF 2016, Bloomington, IN, USA, June 6–10, 2016

T. Arkhangelskiy, N. Serdobolskaya, M. Usacheva. Corpus-oriented lexicographic database for Beserman Udmurt. In: Acta Linguistica Academica, Vol. 64, No. 3, 2017, P. 397-415.

Т. Архангельский. Выявление диалектных особенностей удмуртского языка при помощи интернет-корпуса // VI международная молодёжная научно-практическая конференция Института социально-гуманитарных наук Тюменского государственного университета “Множественность интерпретаций: цифровая перезагрузка”. Тюмень, 14–17 февраля 2018

О проекте создания факсимильно-транскрипционного корпуса рукописей Пушкина

Перцов Николай Викторович
ИРЯ РАН
(Москва)

Факсимильно-транскрипционный корпус рукописей (ФТК; иначе – рукописный факсимильно-транскрипционный корпус) – это корпус, единицами которого служат факсимильно-транскрипционные представления (ФТП), т.е. пары вида «страница рукописи, транскрипция этой страницы». Кроме этих необходимых компонентов в составе ФТП могут быть даны и другие, «факультативные», «поля», часто представленные в архивных описаниях или в публикациях писателей: характер текстов (черновой, беловой с поправками, беловой), отнесение текста к тому или другому произведению, особенности бумаги, пишущих средств, пометы посторонних лиц, идентификация рисунков, датировка, послойное представление вариантов текста и другие сведения.

Будет охарактеризован проект создания ФТК рукописей Пушкина и рассказано о проведённой работе по транскрибированию и компьютеризации его рукописей. В пушкинской текстологии накоплено довольно обширное множество транскрипций черновиков, разбросанных в многочисленных работах. Прежде всего, желательно собрать и «компьютеризовать» эти транскрипции (проведя по возможности их проверку).

Пушкинское рукописное наследие насчитывает примерно двенадцать с половиной тысяч страниц. Это большая величина, но всё-таки отнюдь не астрономическая. Полное транскрибирование Пушкинских рукописей, при участии в нем центральной рабочей группы исследователей и при достаточном финансировании этой работы можно осуществить за десять-двадцать лет, а то и быстрее. Такой срок вполне сопоставим со сроками выхода Большого академического шестнадцатитомника, охватившего период в пятнадцать с небольшим лет (если не считать справочного семнадцатого тома, вышедшего в 1959 году).

К такому электронному факсимльно-транскрипционному представлению и изданию рукописей должны быть привлечены не только силы текстологов – разумеется, требуется участие компьютерных специалистов.

Проект интерактивного словаря компаративных тропов русской литературы XIX-XXI вв.

Петрова Зоя Юрьевна, Ребецкая Наталия Александровна, Фатеева Наталья Александровна

Институт русского языка им. В. В. Виноградова РАН
(Москва)

Проект направлен на создание лексикографического продукта нового типа в рамках междисциплинарного взаимодействия компьютерной лингвистики и авторской лексикографии. Он решает проблему общедоступного представления накопленного обширного языкового материала, описывающего метафорическую картину мира русской литературы, в удобном для пользователя электронном варианте. Интерактивный словарь позволит решить задачу многомерного поиска и структурирования информации о компаративных тропах в русской литературе и создать общую платформу для дальнейшей автоматизации поиска по другим словарным материалам.

Интерактивный словарь будет создаваться на основе пяти выпусков «Материалов к словарю метафор и сравнений русской литературы XIX-XX вв.» [Кожевникова, Петрова 2000; 2010; 2015; 2017]. В «Материалах к словарю…» отражена система метафор и

сравнений русской литературы XIX-XXI вв. и эволюция этой системы. Основные параметры описания компаративных тропов в словаре – семантический инвариант, формальная конструкция и временна́я характеристика. Четкая и единообразная структура словарных статей всех пяти выпусков и единая семантическая классификация тропов делают возможным разработку на основе словарных материалов единой электронной базы данных, которая, в свою очередь, станет основой информационно-поискового аппарата предлагаемого интерактивного словаря.

Будет разработано приложение (программный продукт), способное работать на локальном компьютере или через Интернет. Приложение будет состоять из двух частей: интерфейс пользователя и база данных. В случае приложения для настольного компьютера база данных поставляется вместе с программой на диске. В случае работы через Интернет база данных размещается на сервере в Интернете, а приложение представляет собой веб-страницу.

Взаимодействие пользователя с контентом будет реализовываться через разные типы запросов. Структура базы данных позволит осуществлять поиск по следующим основным элементам словарной статьи: предмет сравнения (что сравнивается), образ сравнения (с чем сравнивается), семантический класс предметов сравнения и образов сравнения, тип конструкции (метафора, сравнение, формальные подклассы сравнений: с глаголами, с прилагательными и т.п.), автор, год.

Проектируемый словарь станет гибким, наглядным инструментом изучения метафорической картины мира русской литературы, характеризуя компаративные тропы во многих аспектах и позволяя быстро получать ответы на самые разные пользовательские запросы.

Список литературы

Кожевникова Н.А., Петрова З.Ю. Материалы к словарю метафор и сравнений русской литературы XIX-XX вв. Выпуск 1 «Птицы». М.: «Языки русской культуры», 2000.

Кожевникова Н.А., Петрова З.Ю. Материалы к словарю метафор и сравнений русской литературы XIX-XX вв. Выпуск 2 «Звери, насекомые, рыбы, змеи». М.: «Языки славянских культур», 2010.

Кожевникова Н.А., Петрова З.Ю. Материалы к словарю метафор и сравнений русской литературы XIX-XX вв. Выпуск 3 «Растения». М.: «Языки славянских культур», 2015.

Кожевникова Н.А., Петрова З.Ю. Материалы к словарю метафор и сравнений русской литературы XIX-XX вв. Выпуск 4 «Камни, металлы», Выпуск 5 «Ткани, изделия из тканей». М.: Издательский дом ЯСК, 2017.

Создание корпуса русской речи носителей автохтонных языков Севера Сибири и Дальнего Востока[15]

Плешак Полина Сергеевна
МГУ, ИЯз РАН
(Москва)

Стойнова Наталья Марковна
ИРЯ РАН, ИЯз РАН
(Москва)

Хомченкова Ирина Андреевна
МГУ, ИЯз РАН
(Москва)

В докладе будет представлен проект по созданию корпуса русской речи носителей автохтонных языков Севера Сибири и Дальнего Востока. Работа ведется в рамках более общего проекта «Динамика языковых контактов в циркумполярном регионе» (http://iling-ran.ru/main/departments/typol_compar/circumpolar). Корпус представляет собой аннотированную коллекцию контактно-обусловленной устной речи билингвов-носителей автохтонных языков указанного региона: самодийских, тунгусо-маньчжурских, чукотско-камчатских.

В коллекцию вошли тексты, собранные как «побочный продукт» экспедиций по документации соответствующих автохтонных языков. Это спонтанные устные тексты, в основном короткие нарративы: фольклор, биографии, этнографические описания, в некотором объеме бытовые диалоги. Прежде всего, это переводы / пересказы текстов на автохтонном языке или тексты, рассказанные на русском языке вместо ожидаемого автохтонного.

Всего в аудиоколлекции сейчас ок. 90 часов русских текстов от носителей нганасанского, разных диалектов ненецкого, лесного и тундрового энецкого, нанайского, ульчского, эвенского и чукотского языков. Расшифрована и размечена часть из них: ок. 20 часов. Расшифровка и разметка текстов ведется в программе ELAN, см. Рисунок 1. Небольшая часть

[15] Работа выполнена при поддержке гранта РНФ 17-18-01649 «Динамика языковых контактов в циркумполярном регионе».

размеченных текстов доступна в виде тестового онлайн-корпуса с возможностью поиска по отдельным характеристикам на платформе Tsakonian corpus platform (Tsakorpus), см. Рисунки 2а, 2б. К осени планируется выложить в онлайн-корпус весь объем расшифрованных текстов с возможностью поиска по всей разметке, выполненной в ELAN, по метаразметке, а также по грамматическим характеристикам.

Тексты расшифровываются в стандартной орфографии с разбивкой на ЭДЕ (≈клаузы) и минимальным отражением интонации. Специальными разработанными в ходе расшифровки тэгами вручную помечаются случаи отклонений от стандартного русского, имеющих очевидную или потенциальную контактную природу. Отдельно (в разных «слоях» разметки) отмечаются нестандартные особенности в области:

- синтаксиса (нестандартная аргументная структура, разного рода рассогласования, нестандартный порядок слов и под.);
- морфологии (нестандартный выбор вида, времени, числа и под., нестандартные формы);
- лексики (материальные заимствования и кальки);
- фонетики (отмечаются только особенно яркие случаи фонетической интерференции);
- прочие (особенности просодии, нестандартное употребление дискурсивных маркеров и др.).

На каждую область заведено от 3 до 20 тэгов. Каждый тэг (в т.ч. морфологические, фонетические) условно привязан к клаузе (в которой встретилось соответствующее явление). На одну клаузу может приходиться любое количество тэгов одного слоя разметки. Используемый формат разметки не предусматривает иерархической структуры тэгов внутри одного слоя (так, например, для нестандартного согласования по роду глагола и прилагательного заведено два отдельных тэга уровня «синтаксис», а не тэг «рассогласование» с двумя подтипами). Отдельно (и менее последовательно) отмечаются яркие отклонения от стандартного русского неконтактной природы (диалектные, региональные особенности).

Имеется также метаразметка, отдельно по рассказчикам (см. Рисунок 3), отдельно по текстам (см. Рисунок 4). Метаразметка по рассказчикам содержит, в частности, базовую социолингвистическую

информацию: год рождения, пол, место рождения и проживания, уровень образования, возраст овладения русским, примерный уровень владения автохтонным языком (языками). Метаразметка по текстам содержит техническую информацию о записи, расшифровке, объеме текста, а также тип и жанр текста.

Создаваемый в рамках этого проекта корпус планируется в дальнейшем расширить, пополнив текстами на контактно-обусловленном русском языке носителей других языков России.

На основе корпуса уже начат ряд исследований по отдельным контактным явлениям в русском языке билингвов: опущению предлогов, рассогласованию по роду, структуре посессивной группы, нестандартным стратегиям сочинения, интерференции в области просодии и др.

Рисунок 1. Фрагмент расшифровки и разметки в программе ELAN

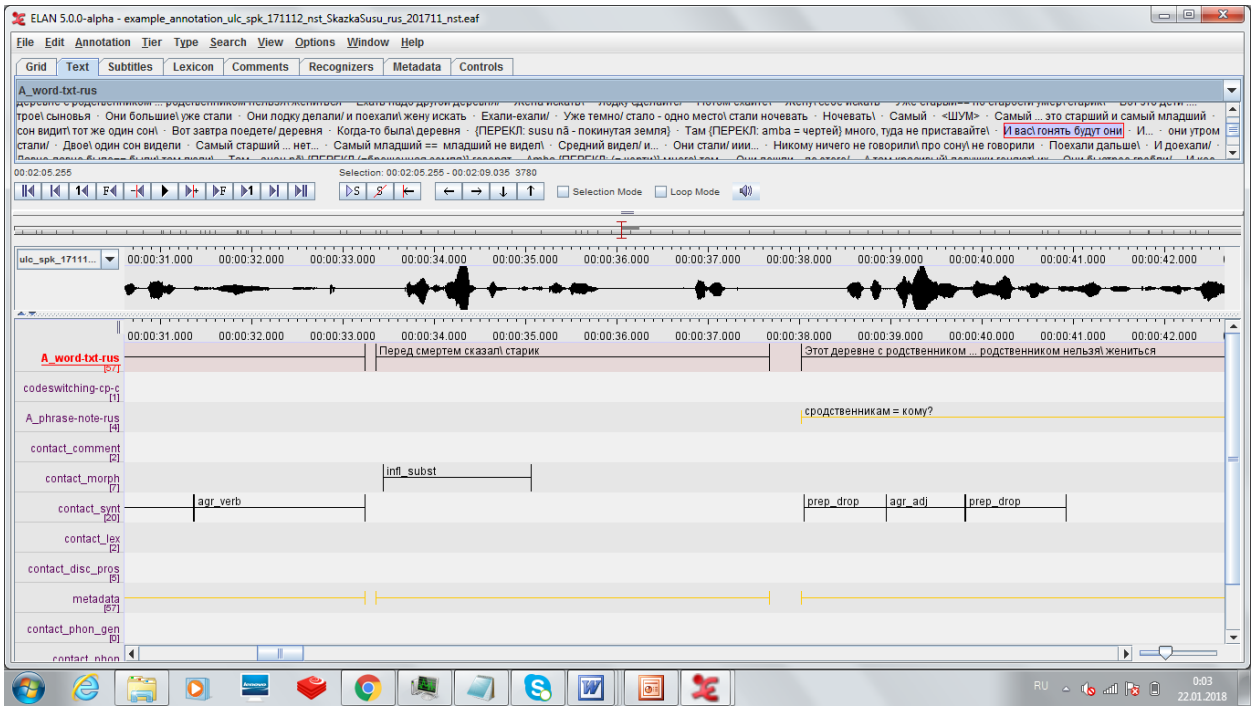


Рисунок 2а. Строка поискового запроса в онлайн-корпусе

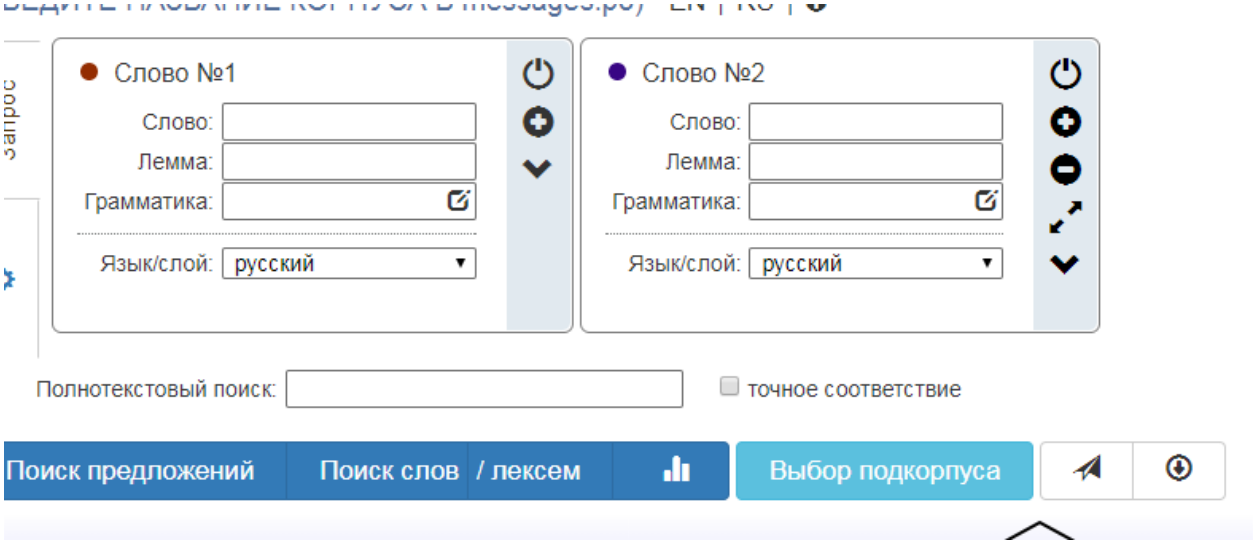


Рисунок 2б. Выдача примеров в онлайн-корпусе

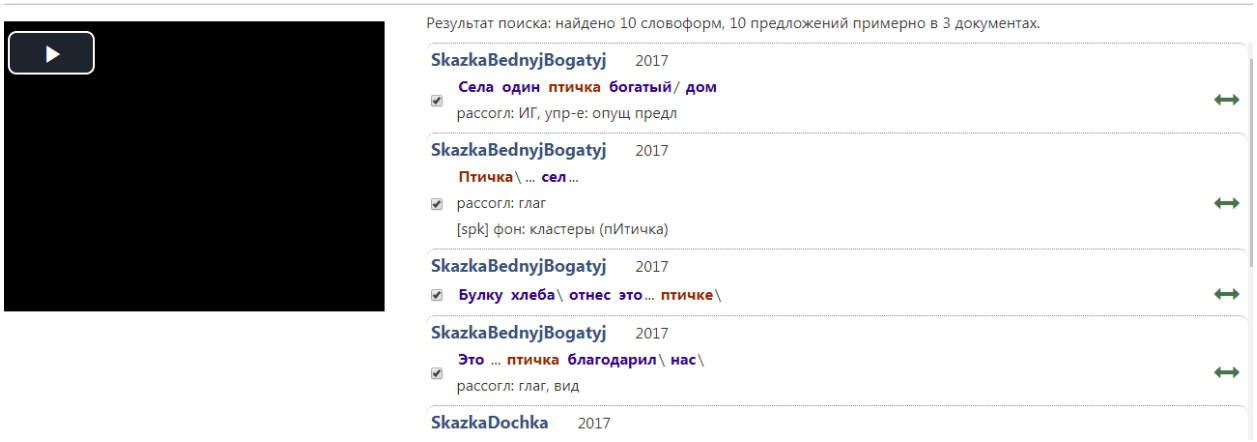


Рисунок 3. Метаразметка по рассказчикам

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	код	L1	год рожд	место ро	место жи	русский (	русский (	возраст с	другие яз	L1 (1...10)	L1 (комме	социолингв	образова
2	vsg	нанайский	1932	Кондон (С	Харпичан	1	очень нес	школа		10		в детстве го	начальное
3	fna	нанайский	1930	Дада	Даерга	1	очень нес	школа		10	тексты	в детстве го	начальное
4	oab	ульчский	1935	Дуди	Булава	3	достаточн	школа	?нанайски	10	тексты	в Булаве с 1	начальное
5	spk	ульчский	1930	Удан	Булава	2	достаточн	школа		10	тексты	родилась в	начальное

Рисунок 4. Метаразметка по текстам

название	длительность	L1	L1 код	код информанта	записал	дата записи	место записи	содержание	версия на L1	русский (оценка стандартности)	русский (комментарий)	расшифровано	расшифровал	дата расшифровки	размечено	размечал	тип текста	жанр
gld_vsg_110813_so_SonZoloto&SonDocheri_rus.wav	0:04:50	нанайский	gld	vsg	Н. Стойнова	110813	Харпичан	как дочь ВС видела во сне золотой слиток, который до этого видела во сне сама ВС		1		да	Н. Стойнова	201710	да	Н. Стойнова	нарратив	случай
gld_vsg_110813_so_DetskijeBerestjanyjeGroby_rus.wav	0:03:44	нанайский	gld	vsg	Н. Стойнова	110813	Харпичан	про детское кладбище и гробы из бересты	vsg_110813_ns_DetskijeBerestjanyjeGroby.wav	1		да	Н. Стойнова	201710	да	Н. Стойнова	описание	этнография
								про то, как у нанайцев нельзя было рожать в доме, строили										

Национальный корпус русского языка: история проекта и некоторые результаты

Плунгян В.А.
ИРЯ РАН
(Москва)

Проект Национального корпуса русского языка в разных отношениях тесно связан с Машинным фондом русского языка, в создание которого В. М. Андрющенко вложил столько сил и энергии. В докладе дается краткий обзор истории корпусного проекта, как на начальном этапе (замыслов и обсуждений), так и на современном этапе (в связи с проблемами развития и ещё не решёнными задачи). Обсуждается возросшая роль корпусных методов (и "корпусной идеологии" в целом) в современной лингвистике.

Словари в справочно-информационной системе СКАЗКА-2

Рафаева Анна Валерьевна
НИВЦ МГУ им. М.В. Ломоносова
(Москва)

Справочно-информационная система СКАЗКА-2, разрабатываемая автором, является работающим прототипом АРМ фольклориста и служит для исследования языка и сюжетной структуры русской волшебной сказки. В систему входят следующие компоненты:

- Корпус сказочных текстов по ряду авторитетных сказочных сборников (1,3 млн. словоупотреблений). Корпус продолжает пополняться. Все исследования сначала проводятся на материале сборника А.Н. Афанасьева, после чего полученные выводы проверяются и уточняются по другим сборникам, входящим в корпус.
- Программные средства для работы с текстом (составление частотного словаря, составление конкордансов по заданным ключевым и стоп-словам, простые подсчеты).
- Программные средства для представления и визуализации полученных данных. Используется либо свободно распространяемое ПО, либо ПО, разработанное автором. Так, для графического представления хранения заданных пользователем отношений между объектами служит разработанное автором программное обеспечение, в то время, как графическое представление этих связей осуществляется с помощью пакета GraphViz. В качестве объекта может выступать сказочный персонаж, локус или роль персонажа по В.Я. Проппу; отношения между ними задаются пользователем в зависимости от поставленной задачи, например, *состоять в родстве*.
- Электронные версии ряда толковых словарей и ряд простых программ для извлечения данных из словарных статей.
- Частотный словарь словоформ, создаваемый программно.
- Материалы к словарю сказочных персонажей.

Именно словарям, их использованию в алгоритмах автоматизированного анализа и разработке собственных внутрисистемных словарей, и будет посвящен доклад.

Литература

Пропп В. Я. Морфология <волшебной> сказки. Исторические корни волшебной сказки. – М.: Лабиринт, 1998.

Народные русские сказки А. Н. Афанасьева: В 3 т. / Подгот. Л. Г. Бараг, Н. В. Новиков; Отв. ред. Э. В. Померанцева, К. В.Чистов. — М., 1984–1985.

Reword. Бесплатная программа-словарь. – Режим доступа: http://reword.org/online

Словари Онлайн. http://slovarionline.ru/malyiy_akademicheskiy_slovar

Даль В. И. Толковый словарь живого великорусского языка : в 4 т. / В. И. Даль. – 4-е изд., стереотип. – М., 2007.

Словарь русского языка: В 4-х т. / РАН, Ин-т лингвистич. исследований; Под ред. А. П. Евгеньевой. — 4-е изд., стер. — М., 1999. Режим доступа: http://feb-web.ru/feb/ushakov/ush-abc/default.asp

Ожегов С. И., Шведова Н. Ю. Толковый словарь русского языка: 80 000 слов и фразеологических выражений. — 4-е изд., М., 1997. — 944 с.

Толковый словарь русского языка: В 4 т./ Под ред. Д. Н. Ушакова. — М., 1935—1940. – Режим доступа: http://feb-web.ru/feb/ushakov/ush-abc/default.asp

Малые корпуса: проект НИУ ВШЭ

Екатерина Владимировна Рахилина,
НИУ ВШЭ, ИРЯ РАН
(Москва)

Принято считать, что корпуса как инструмент исследования всегда оперируют так называемыми большими данными (big data) – и действительно, в современном исполнении полнотекстовые базы по развитым литературным языкам представляют сотни и сотни миллионов

словооупотреблений. В то же время, есть задачи, для которых такие объемы текстов недостижимы даже и для мощных литературных языков, однако корпусные методы им необходимы и хорошо на них работают. Речь идет об учебных корпусах (Learner corpora). Это корпуса текстов с явными отклонениями против «канонического» литературного языка, прототипически, написанные не-носителями, изучающими данный язык. Как правило, современные коллекции такого рода текстов не очень большие (если не считать промышленных, в которых системно собираются экзаменационные задания), но для них разрабатывается специальная разметка, систематизирующая языковые отклонения. Исследования с опорой на статистические данные по таким корпусам позволяют выявить закономерности в возникновении и динамике отклонений.

Школа лингвистики НИУ ВШЭ создала несколько учебных корпусов языка.

Во-первых, это Учебный корпус русского языка (RLC) – это корпус русского как неродного, включающий не только тексты изучающих русский иностранцев из разных стран и с разным доминантным языком, но и тексты сбалансированных и несбалансированных билингвов (в другой терминологии – эритажных носителей). В нем в сотрудничестве с лингвистами и преподавателями разных стран собираются письменные тексты от носителей разных доминантных языков -- прежде всего, европейских (английского, немецкого, финского, итальянского…) – но также, например, японского или казахского. Они автоматически размечены морфологически и вручную – по небольшому набору метапризнаков, п также по специально разработанным признакам, существенным при изучении отклонений от стандартной нормы – в основном, морфологическим, синтаксическим и лексическим. В докладе будут показаны примеры задач, которые с помощью этого ресурса можно решать.

Другой крупный ресурс – это Корпус русских учебных текстов (КРУТ). В нем собраны студенческие работ русскоязычных авторов (курсовые и дипломные работы, эссе и проч.), нуждающиеся в правке, часто ввиду интерференции с английским. (Есть и его аналог -- Учебный корпус английского языка (REALEC), представляющий собой корпус сочинений русскоязычных студентов, написанных по-английски – и тоже со случаями интерференции).

Кроме классических учебных корпусов в работе находится еще несколько корпусных ресурсов, также фиксирующих отклонения от современного литературного языка. К ним относится корпус языка XIX века (прежде всего, детально размеченный корпус языка «Героя нашего времени» М.Ю. Лермонтова), корпус детских письменных работ, а также устный диалектный корпус – корпус устяньского языка. Каждый из них имеет свои особенности и создан для решения своих задач, но в целом этот языковой материал является для русского языка источником данных о природе и характере языковой интерференции с другими языками, а также о потенциальных или совершившихся языковых изменениях в самом русском.

В докладе будет дан общий обзор этих корпусов, работа с некоторыми из них (RLC корпусом языка XIX века) будет рассмотрена подробно на примерах.

Проект словаря языка А. П. Чехова

Н.А. Ребецкая
Институт русского языка имени В. В. Виноградова РАН
(Москва)

Ставится задача создать авторский словарь нового типа, сочетающий в себе традиционное бумажное представление результатов с электронной базой данных, дружественной к пользователю словаря.

К настоящему времени создан конкорданс ко всем художественным произведениям А. П. Чехова с группировкой контекстов по разным значениям для каждой лексемы (на основе толкового словаря С. И. Ожегова). Конкорданс конвертируется в базу данных, таблицы которой включают несколько полей. В программной оболочке C++ Builder создана программа работы с базой данных. Она позволяет последовательно просматривать записи, расширять контексты, осуществлять выборку, в том числе и словообразовательных гнезд, сохранять массивы в текстовом и RTF – формате, просматривать полный текст произведения.

Далее представлен фрагмент работы программы (рис.1):

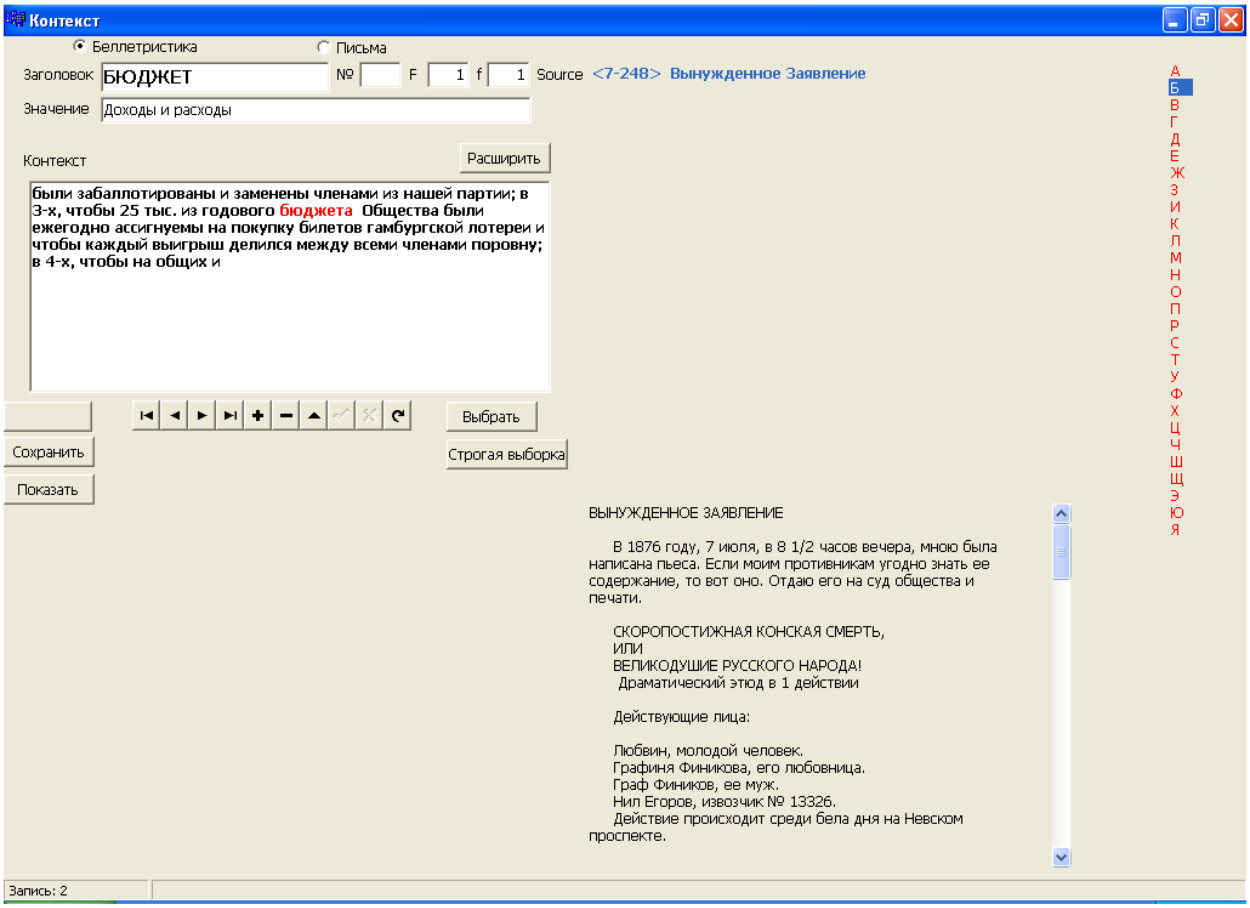


Рис. 1. Форма работы с БД на примере словарной статьи «БЮДЖЕТ»

На форме можно увидеть следующие компоненты:

Поле заголовка, номер значения, общая частота, частота значения, адрес фрагмента (он отображен в следующем формате: номер тома, номер страницы начала произведения, заглавие), поле семантизации, контекст. Компонент “навигатор”, расположенный под контекстом, обеспечивает перемещение указателя текущей записи к следующей, предыдущей, первой и последней записи. Справа внизу - окно, отображающее полный текст произведения. Его можно открыть щелчком на компоненте Source.

При необходимости контекст можно расширить (рис.2).

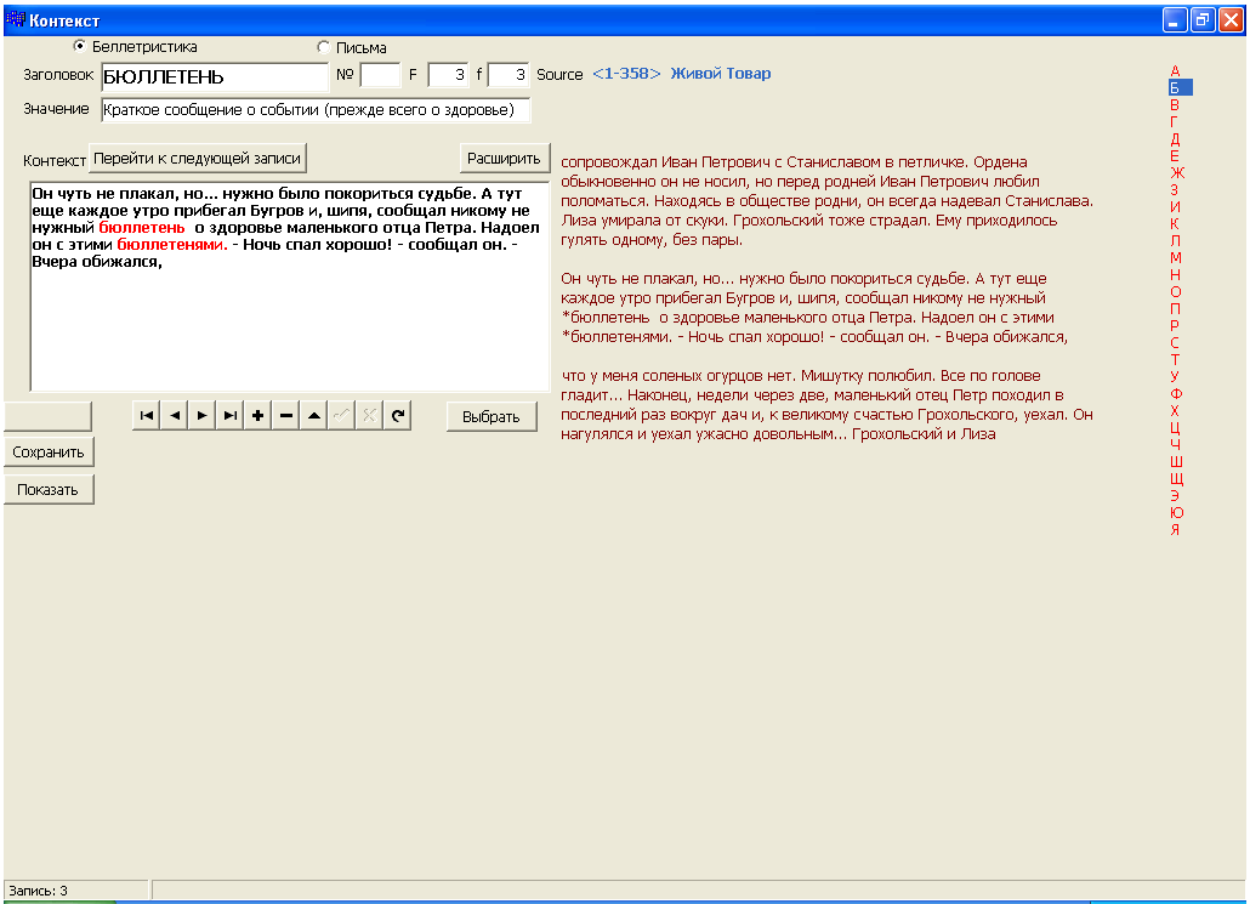


Рис. 2 Словарная статья «БЮЛЛЕТЕНЬ» с расширенным контекстом

Компоненты **Выбрать** и **Показать** позволяют выбрать нужные записи, в том числе по шаблону, и вывести их на форму в виде списка.

Ниже приведен список записей выборки с поисковым элементом «БЮСТ»:

Лексема Общая частота № значения Показ значения Част. знач.

БЮСТ 6 1 Скульптура 3

Бюсты и портреты знаменитых писателей глядят на его быстро бегающее перо, не шевелятся и, кажется, думают: "Эка, брат, как ты насобачился!"

БЮСТ6 1 Скульптура 3

У меня письменный стол рублей в четыреста, с инкрустациями, бархатная мебель, картины, ковры, *бюсты*, тигровая шкура,

БЮСТ6 1 Скульптура 3

Бюсты Крылова, Пушкина и Гоголя.
Этажерка с чучелами птиц. Шкаф с книгами.

БЮСТ6 2 Женская грудь 3
Красива она не была, но нравиться могла. Лицо было полное, симпатичное, здоровое, а шея, о которой говорил Семен, и *бюст* были великолепны.

БЮСТ6 2 Женская грудь 3
Я не без нечистых мыслей глядел на ее *бюст* и в то же время думал о ней:

БЮСТ6 2 Женская грудь 3
Мысли о загробных потемках не мешали мне отдавать должную дань *бюстам* и ножкам.

БЮСТИК 5 1 Скульптура 2
Бюстики и карточки великих писателей, куча черновых рукописей, том Белинского с загнутой страницей,

БЮСТИК 5 1 Скульптура 2
устроила красивую тесноту из китайских зонтов, мольбертов, разноцветных тряпочек, кинжалов, *бюстиков*, фотографий...

БЮСТИК 5 2 Женская грудь 3
Наружность у нее самая обыкновенная: нос папашин, подбородок мамашин, глаза кошачьи, *бюстик* посредственный.

БЮСТИК 5 2 Женская грудь 3
Прямой носик, дивный *бюстик,* чудные волосы, прелестные глазки - ни одной опечатки! Прокорректировал и женился.

БЮСТИК 5 2 Женская грудь 3
против ее улыбки, против неги, которою так и дышит ее миниатюрный, словно выточенный *бюстик*.

Всего 11.

Поисковая система предполагает также поиск по словообразовательному гнезду. Для удобства пользователя создается список всех таких гнезд (в соответствии со Словообразовательным словарем русского языка А.Н. Тихонова.)

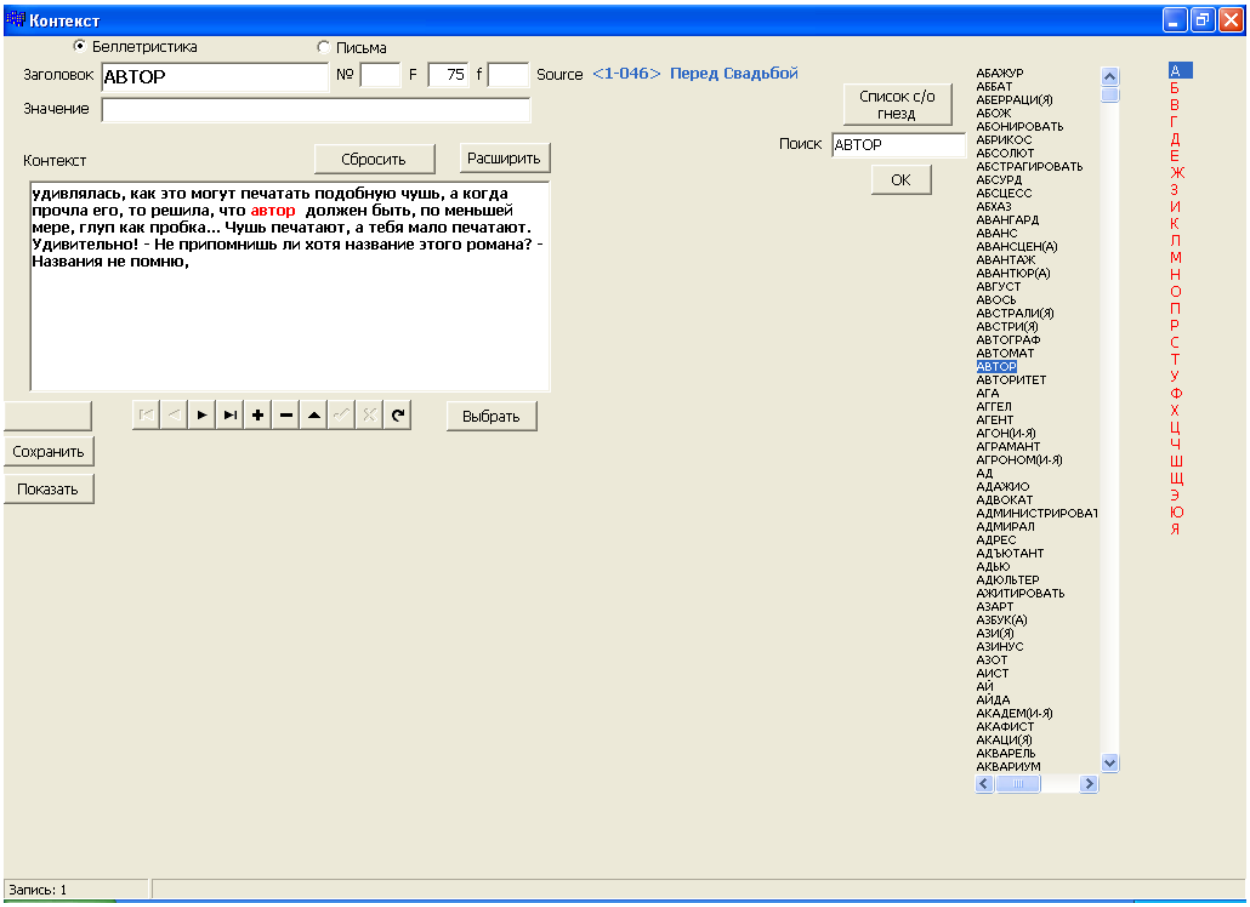


Рис. 3 Форма представления словарной статьи «АВТОР». Справа список словообразовательных гнезд

Выбрав из этого списка элемент «АВТОР», получим словарные статьи для слов АВТОР, АВТОРСКИЙ, АВТОРСТВО, АВТОРША. Поиск по шаблону «АВТОР» покажет дополнительно слова АВТОРИТЕТ, АВТОРИТЕТНЫЙ, не входящие в данное гнездо.

Особенно важен поиск по списку словообразовательных гнезд в случае чередующихся корней. Например, выбор элемента «МЕСТИ» отобразит словарные статьи с заголовками МЕСТИ, МЕТЕЛЬ, ВЫМЕСТИ, МЕТЕЛИЦА, МЕТЛА, ПОДМЕСТИ, ПОМЕЛО, а элемент «МОЛОТЬ» произведет ЗАМОЛОТЬ, МУКОМОЛЬНЯ, НЕМОЛОТЫЙ, ПОМОЛ, МЕЛЬНИК, МЕЛЬНИЦА, МЕЛЬНИЧНЫЙ, МЕЛЬНИЧКА, ПЕРЕМОЛОТЬСЯ, ПУСТОМЕЛЯ.

Задача составителей - создать авторский словарь нового типа, характерные черты которого - исчерпывающее и легко доступное представление многозначности для всех знаменательных слов; гибкость в выборе пользователем интересующих его подкорпусов (в т. ч. нетрадиционных), а также лингвистических объектов (слов в традиционном смысле, словообразовательных гнезд).

Кроме того, к традиционному расщеплению полисемии добавляются такие аспекты, как статистический словарь языка писателя с подразделением на подкорпусы, словарь текстуальных связей, перечень лексических маркеров Чехова на фоне современной ему прозы. Тем самым должен быть наведен мост между формальной статистической лексикографией и лексикографией традиционного типа.

Идея машинного фонда языка в контексте кардинальных технологических изменений

Учреждение образования «Гродненский государственный университет имени Янки Купалы»
rychkova@grsu.by

Известно, что в разработке концепции и популяризации идеи машинного фонда русского языка (далее – МФРЯ), выдвинутой академиком А.П. Ершовым, неоценимую роль сыграл В.М. Андрющенко, который лично встречался не только с учеными, но и со студентами отделений математической лингвистики, многие из которых затем, став специалистами, участвовали в работах, связанных с МФРЯ. Сегодня мало кто знает, что работы по созданию ресурсов МФРЯ велись во многих республиках Советского Союза, в том числе и в Гродненском государственном университете имени Янки Купалы в Белоруссии под руководством Г.В. Ермоленко.

После распада Советского Союза идея машинного фонда продолжала жить и реализовывалась посредством разработки банков ресурсов национальных языков. Так, в Республике Беларусь в рамках Программы развития средств электронных коммуникаций в Украине, Беларуси и Молдове, финансируемой фондом Евразия, в 1996-1997 гг. под

руководством профессора Г.А. Цыхуна был реализован проект «Электронный фонд белорусского языка и сеть лингвистической информации в Беларуси», объединивший усилия ученых Национальной академии наук и высших учебных заведений, членов Терминологической комиссии при Министерстве образования и науки Республики Беларусь и ряда общественных организаций. Научный совет проекта, в который, помимо Г.А. Цыхуна, входило еще семь ученых, в том числе и автор данного доклада, сумел эффективно организовать работу по оцифровке разнообразных лексикографических продуктов, осуществив, по сути, инвентаризацию лексических ресурсов белорусского языка, включая терминологические ресурсы. Была также создана библиографическая база данных по языкознанию и электронный массив текстов научных публикаций на белорусском языке. К сожалению, ресурсы электронного фонда постигла та же участь, что и ресурсы МФРЯ.

Кардинальные технологические изменения последних десятилетий позволяют по-новому взглянуть на идею МФРЯ как идею экспансии «русского мира» в интернет-пространстве. На примере развития лингворесурсологии в Гродненском государственном университете имени Янки Купалы в докладе будут показаны возможности сотрудничества в реализации проектов по созданию нетрадиционных электронных ресурсов русского языка и двуязычных (русского и национального языков) ресурсов, а также необходимость использования подобных ресурсов в подготовке специалистов по прикладной лингвистике. Обосновывается жизненность идеи машинных фондов языков и доказывается необходимость возрождения концепции машинного фонда русского языка с учетом возможностей пространства ‘открытой науки’.

Машинный фонд и национальный корпус русского языка: преемственность и новации[1]

Савчук Светлана Олеговна
Институт русского языка им. В.В. Виноградова РАН
(Москва)

Идея создания Машинного фонда русского языка, впервые сформулированная А.П. Ершовым в 1978 году, получила горячую поддержку и после коллективного обсуждения к середине 1980-х годов обрела черты конкретного проекта. Сегодня, более тридцати лет спустя, можно сказать, что проект состоялся, хотя и претерпел значительные изменения вследствие произошедших за эти годы общественных и технологических трансформаций. Задача «комплексной автоматизации лингвистических исследований и прикладных разработок» [Машинный фонд: 234] оказалась своевременной и плодотворной, и начало работы над ее реализацией дало толчок к развитию отечественной компьютерной и корпусной лингвистики. В настоящее время создано немало электронных ресурсов, которые успешно используются в научных филологических исследованиях и преподавании русского языка. К ним относятся электронные библиотеки, лингвистические корпусы текстов, ресурсы на основе корпусов – корпусные грамматики и электронные словари, учебно-методические порталы.

Среди наиболее известных корпусов русского языка можно назвать Национальный корпус русского языка (http://ruscorpora.ru/), Генеральный интернет-корпус русского языка (http://www.webcorpora.ru/), интернет-корпус русского языка, созданный в Словацкой академии наук – Araneum Russicum (http://sketch.juls.savba.sk/aranea_about/_russicum.html), корпусы звучащей речи (http://spokencorpora.ru/). Как представляется, концепция наполнения и организации текстового модуля Машинного фонда в наибольшей степени реализована в проекте Национального корпуса русского языка. НКРЯ можно считать современным воплощением и одновременно развитием идей Машинного фонда русского языка.

В докладе будет рассмотрена общая архитектура НКРЯ, которая складывается как из отдельных корпусов, так и из текстовых коллекций внутри одного большого корпуса. Критерием для выделения класса текстов в самостоятельный корпус является наличие особой лингвистической разметки и соответствующих особых параметров поиска. На этом основании отдельными корпусами являются параллельный, поэтический, газетный, синтаксический, диалектный, обучающий корпусы, модуль корпусов устной речи (устный, акцентологический, мультимедийный, параллельный мультимедийный), модуль исторических корпусов (древнерусский, старорусский, церковнославянский) и др. И напротив, однородность разметки позволяет объединить в составе одного корпуса множество разнородных текстов, относящихся к разным функциональным сферам, – публицистической, научной, художественной, официально-деловой, церковно-богословской, рекламной, обиходно-бытовой, электронной коммуникации. Сложная система метаописания текстов по многим параметрам и наличие интерфейса дает возможность выбора пользовательского подкорпуса по каждому из значений этих параметров или их комбинации. В докладе будет дан анализ состава и структуры основного корпуса письменных текстов в сопоставлении с проектом Машинного фонда и с точки зрения соответствия критериям репрезентативности и полноты, а также намечены пути его развития.

Национальный корпус русского языка – это живой, открытый, развивающийся проект с четко обозначенными ближайшими целями и перспективными разработками методов и инструментов обработки, накопления и представления данных.

Литература

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

_____

[1]Работа выполнена при поддержке программы Президиума РАН «Памятники материальной и духовной культуры в современной информационной среде».

К модернизации прикладного семантического словаря: анализ употребления метаязыковых единиц

Семенова Софья Юльевна
ИНИОН РАН, РГГУ
(Москва)

Одним из направлений деятельности Машинного фонда русского языка, актуальных и ныне, стало накопление словарных ресурсов, в т. ч., ориентированных на прикладные задачи. К таким ресурсам может быть отнесен, в частности, русский семантический словарь РУСЛАН для АОТ. Первые версии РУСЛАНа (расшифровывается как «РУсский СЛоварь для АНализа») были получены на рубеже 1990-х—2000-х гг. под руководством Н.Н. Леонтьевой. Разработки начались в Институте США и Канады РАН, а затем были перебазированы в НИВЦ МГУ.

Структуру статьи и метаязык разработала Н.Н. Леонтьева. Фактически, структура и метаязык вобрали многолетний опыт, накопленный Н.Н.Леонтьевой в компьютерной лингвистике. Прообразом РУСЛАНа был словарь системы ФРАП (французско-русского автоматического перевода), разрабатывавшейся под ее руководством в 1970-е — 80-е гг.

Над составлением статей РУСЛАНа в разные годы наряду с Н.Н. Леонтьевой и автором работали Е.В. Горелик, М.В. Ермаков, А.С. Панина, М.С. Шаталова, О.А. Штернова и некоторые другие специалисты.

В конце 2000-х гг. работы над словарем были приостановлены из-за проблем финансирования. С 2017г. был выделен грант РФФИ (РГНФ) на модернизацию словарной системы (проект РФФИ № 17-04-00594-ОГН «Автоматический словарь РУСЛАН: обновленная концепция, новая лексика»). Сейчас, с одобрения Н.Н. Леонтьевой, над словарем вместе с автором (руководителем проекта) работают М.В. Ермаков, С.А. Крылов, А.С. Панина.

Словарь содержит сведения о семантических классах слова, валентной структуре, лексической сочетаемости, фразеологии, тезаурусных связях, энциклопедических функциях, эмпирическом информационном весе, отраслевой принадлежности, английских эквивалентах. Для лексем полисемичных слов предусмотрены отдельные словарные статьи.

Модернизация предполагает серию работ, в т.ч. редактирование словаря по его полям, некоторое развитие формата, расширение словника.

«Старый» РУСЛАН насчитывал около 12000 входов; количество статей необходимо увеличить. Расширение осуществляется, в основном, по статистическим критериям — заполняются статистические лакуны в словнике.

В «новом» РУСЛАНе (т.е. создающемся с 2017г.) расширена иллюстративная зона. Наряду с модельными примерами в духе в «старого» РУСЛАНа, привлечены примеры из НКРЯ. Зона иллюстраций теперь включает иллюстрации общего характера, «прицельные» иллюстрации по валентностям, иллюстрации фразеологических единиц и поле особых случаев, где собираются примеры для последующих детализаций.

Массив «правых частей», являясь коллективным продуктом, в определенной мере отражает возможные трактовки единиц метаязыка. Принципиальными представляются прежде всего те (абстрактные) единицы, что употребляются для кодирования семантики слова-заголовка и его участников.

Семантика отражается в форме конъюнкций дескрипторов трех типов:

1) унарных семантических характеристик (СХ: АБСТРактное»; ВМЕСТЛ - вместилище» и др.);

2) бинарных смысловых отношений (СО: ИДЕНТ (А, В) – «А идентификатор В»; ИСТОЧник (А, В); СО предусмотрены для именования валентностей, но используются и как унарные СХ);

3) некоторых лексических функций.

Конъюнкции фактически моделируют компонентную структуру слова.

(Также в «новом» РУСЛАНе для отражения смысла ряда предикатных слов апробируется дополнительный аппарат: смысл представляется последовательностью элементарных ситуаций; такие описания, основываясь на идеях Н.Н. Леонтьевой, строит М.В.Ермаков.)

Наблюдаемый выбор метаязыковых единиц составляет определенную тему для изучения. Так, разработчик программной среды «старого» РУСЛАНа, А.В. Сокирко, еще в 1998г. проделал компьютерное исследование употреблений СХ ИНФОРМация и НОС_ИНФ («носитель информации»).

Предполагается, что в сообщении будут приведены данные об использовании ряда других СХ и СО: ФУНКЦия, СОДЕРЖание, СПЕЦиализация, СВЯЗАН и др.: распределение по основным семантическим группам заглавных слов; соотношение переносных и буквальных

пониманий; некоторые эмпирические правила разделения близких концептов; для некоторых дескрипторов — сложившиеся позиции в конъюнкциях.

Анализ реального употребления дескрипторов обусловлен практической задачей выравнивания словарных описаний. Он представляется полезным и с теоретической точки зрения, так как способствует уточнению внутреннего содержания концептов, обозначаемых метаязыковыми единицами, и показывает экстенсионал таких единиц.

Сведения об авторе: Семенова Софья Юльевна, доклад «К модернизации прикладного семантического словаря: анализ употребления метаязыковых единиц», старший научный сотрудник ИНИОН РАН, доцент РГГУ,

e-mail: sonya_sem@mail.ru

Лексико-семантическое поле цвета в поэзии Андрея Белого

Тарумова Наталья Тимофеевна
НИВЦ МГУ им. М.В. Ломоносова
(Москва)

Доклад посвящен выделению и структурно-семантическому анализу цветовых номинаций в поэзии Андрея Белого. Цветопись активно используется во всех жанрах литературы как яркое и многофункциональное изобразительное средство. Исследователями изучались цветообозначения многих русских поэтов: А. Блока, С. Есенина, В. Маяковского, Ф. Тютчева, М. Цветаевой и др. Объектом нашего анализа явились языковые единицы, содержащие цветовую семантику в поэзии Андрея Белого. Непосредственным материалом исследования стали прижизненные поэтические сборники А. Белого: *Золото в лазури* (1904), *Пепел* (1909 и 1929), *Урна* (1909), *Королевна и рыцари* (1919), *Христос воскрес* (1918), *Звезда* (1922), *Первое свидание* (1921), *После разлуки* (1922), *Стихи о России* (1922), *Сборник Гржебина* (1923), а также использовались архивные материалы К.Н. Бугаевой (жены А. Белого), словники к сборникам стихотворений: *Золото в лазури*, *Пепел*, *Урна*, *Зовы времен.* [1, 2].

При анализе структурно-семантического состава колоризмов в поэзии Андрея Белого были рассмотрены обозначения десяти цветов, считающиеся основными. К числу основных цветонаименований относятся семь названий

цветов радуги, также три ахроматических – серый, белый, черный. [3]. Из этого базового спектра в поэзии Белого представлены все десять наименований цветов, а также авторские цвета и семантические оттенки.

Поэтический словник А. Белого насчитывает 56356 слов из них, 385 лексем имеют семантику цвета с частотой употребления 1658. Колоризмы, слова обозначающие цвет, в лирике Белого представлены практически всеми знаменательными частями речи: прилагательными, существительными, глаголами, наречиями, причастиями, деепричастиями.

В литературном контексте разных эпох и, в частности, символизма конца XIX – начала XX вв. в основном исследуется символикообразующая функция цвета. [4, 5]. В творчестве А. Белого символика цвета обычно анализируется исследователями на трех поэтических сборниках: *Золото в лазури*, *Пепел* и *Урна*. Считается, что именно в этих сборниках Белый наиболее ярко выразил себя как оригинальный поэт, в них более четко воплощена его цветовая поэтика.

В ходе анализа структурно-семантического состава лексем, обозначающих цвет, был выявлен спектр наиболее представленных в поэтических текстах Андрея Белого цветов из числа основных. Самыми частотными оказались желтый (22%), красный (21%), серый (14%).

Авторские семантические оттенки имеют в текстах активно действующие, разнообразно варьирующиеся лексемы – актуализаторы, что является одной из специфических черт поэтики Белого.

Обращение к сложным прилагательным вызывается стремлением избежать грамматического однообразия при нагнетении ряда новых признаков, например: *серебряный, воздушно-серебряный, серебряно-бледный, сереброблещущий, метельносеребристый, метельно-серебряный, сребристожгучий*.

Используя отглагольные цветообозначения, нередко окказионального характера, по грамматическим признакам совпадающие с причастными конструкциями Андрей Белый не просто фиксирует цвет, а передает его действенную силу: *белеющий, огневеющий, огнистоблещущий, пламенеющий, янтареющий, голубеющий, снегоблещущий, золотеющий.*

В тексте встречаются авторские цвета, например: *Одетый в плащ зари вечерне-темный; одуванные берега; свинцовая бледность; желтым бредом; глаза – сплошные синероды.*

Литература

1. Отдел рукописей Российской государственной библиотеки (ОР РГБ, М.). Ф. 25. А. Белый (1880-1934). Т.2.

2. Отдел рукописей Российской национальной библиотеки (ОР РНБ, СПб.). Ф. 60. А. Белый (1880-1934).

3. Василевич А.П., Кузнецова С.Н., Мищенко С.С. Цвет и названия цвета в русском языке. М.: КомКнига, 2005. 216 с.

4. Кожевникова Н.А. Словоупотребление в русской поэзии начала XX века. М.: Наука, 1986. 252 с.

5. Кожевникова Н.А. Язык Андрея Белого. – М.: Институт русского языка РАН, 1992. 256 с.

Компьютерный архив исчезающих языков Восточной Индонезии

Членова Светлана Федоровна
НИВЦ МГУ
(Москва)

Компьютерный архив языков Молуккских островов создан на основе полевых материалов, собранных в 60-х гг. XX в. автором и М.А. Членовым в Восточной Индонезии. С 70-х гг. бесписьменные молуккские языки, по большей части, малочисленные, были целенаправленно вытесняемы государственным индонезийским языком, в XX1 в. процесс этот значительно ускорился в условиях глобализации.

Наши материалы по 31 изолекту (серуа [Членова & Членов 2004: 265-280], дамар [Chlenova 2008: 163-177], аруанские языки [Членов & Членова 2008:192-261], гором [Членова 2010: 360-407], манусела [Членова 2012:128-173], сула-санана [Членова 2017: 197-236] и др.) до сих пор в ряде случаев остаются либо единственными, либо самыми обширными. Их ценность состоит еще и в том, что они документируют синхронный срез восточноиндонезийских языков до периода их ускоренного исчезновения.

Анализ этих материалов лежит в основе наших работ о специфике в молуккских языках грамматических категорий, например, глагола [Членова 2008: 262-289], прилагательного [Chlenova 2012: 309-314] и др.

Три из наших описаний молуккских языков (давлор [Chlenova 2002: 145-175], дамар [Chlenov & Chlenova 2006] и теунский вариант языка ветан [Членова 2008: 262- 289]) являются первыми и пока единственными. Работа по языку дамар включена в международный справочник по языкам мира Ethnologue, а также напечатана в одном из профессиональных справочников,

из серии Icfai's Professional Reference Book (Bangalore: University Press), ориентирующихся на авторитетные статьи, написанные экспертами и опубликованные в ведущих профессиональных научных журналах.

К настоящему времени наши работы по молуккским языкам, содержащие словарные и текстовые данные, выложены в ПДФ версии на сайте ЛАЛС htpp://lcl.srcc.msu.ru и таким образом доступны как для лингвистов, так и для представителей малочисленных этнических общностей, проявляющих острую заинтересованность в материалах по исчезающим языкам.

ЛИТЕРАТУРА

Chlenova, Svetlana F. Daweloor, A Language of Southwestern Moluccas // Малайско-индонезийские исследования, выпуск XV, M 2002. С. 145-175

Членова С. Ф. Членов М. А. Серуа, исчезающий язык в Восточной Индонезии // Малайско-индонезийские исследования, выпуск XVI, M 2004. С. 265-280

Chlenov, Michael A. & Svetlana F. Chlenova. West Damar language or Damar-Batumerah, an isolatate in South-Eastern Indonesia //10 –ICAL, Philippines, январь 2006. Papers (1, 2 п. л.) http://www.sil.org/asia/philippines/ical/papers.html

Членова С. Ф. Теунский вариант языка ветан (Восточная Индонезия): материалы и грамматические заметки// Малайско-индонезийские исследования, выпуск XVII, M. 2006. С. 35-81

Chlenova, Svetlana F. Preliminary grammatical notes on Damar Batumerah (West Damar), a language of Southwest Maluku // Language and text in the Austronesian world (eds Ogloblin A., Lander Yu.), Munchen, LINCOM-Europa. 2008. P. 163-177

Членов М.А. & Членова С.Ф. Аруанские языки: необычная языковая общность в Восточной Индонезии // Малайско-индонезийские исследования, выпуск XVIII, M. Акад. гуманитарных исслед. 2008. С.192-261

Членова С.Ф. Западный таранган и добель: языки активного типа (острова Ару, Восточная Индонезия) // Малайско-индонезийские исследования, выпуск XVIII, M.: Акад. гуманитарных исследований. 2008. С. 262-289

Членова С.Ф. Заметки о языке гором с приложением словника и образцов предложений // Studia Anthropology: Сборник в честь проф. М.А. Членова. Мосты культуры/ Гешарим. Москва-Иерусалим. 2010. С. 360-407

Членова С. Ф. Манусела, язык центрального Серама (Восточная Индонезия): материалы и заметки// Малайско-индонезийские исследования, выпуск XIX, М.: Акад. гуманитарных исследований 2012. С. 128-173

Chlenova, Svetlana F. Category of adjectives in Dawera-Daweloor, a language pf Eastern Indonesia //Языки Дальнего Востока, Юго-Восточной Азии и Западной Африки: Материалы X Международной конференции (Москва, 21-22 ноября 2012), М.: Ключ-С. 2012. С.309-314

Членова С.Ф. Материалы и заметки по сула-санана, языку архипелага Сула в Восточной Индонезии // Малайско-индонезийские исследования, т. XX, М.: ИСАА МГУ им. М. В. Ломоносова, Общество «Нусантра», 2017. С. 197-236

Создание контента интерактивной карты для сайта «Поэзия Московского университета: от Ломоносова и до…»

Шумарина Ирина Викторовна (Shumarina_Irina@rambler.ru)
ЛАЛС НИВЦ МГУ имени М. В. Ломоносова
Москва

В докладе рассматриваются практические аспекты создания контента интерактивной карты для сайта «Поэзия Московского университета: от Ломоносова и до…». В настоящее время вышли в свет восемь книг одноименного интернет-проекта, существующего с 2000 года по адресу http://poesis.ru. Накопленный материал, редкие и архивные документы могут быть использованы для дальнейших литературных и исторических исследований. Не менее важной функцией проекта является популяризация русской поэзии и истории Московского государственного университета. При этом становится важной форма подачи материала.

Для повышения наглядности сайта кандидатом филологических наук Рафаевой А.В. была разработана методика создания интерактивной карты. На карту нанесены все места, в которых жили, учились, служили, были в

отставке или посещали во время путешествий и где ушли в мир иной поэты, связанные с Московским университетом. Имена поэтов были взяты в хронологическом порядке с сайта (со страницы http: //www.poesis.ru/poeti-poezia/fr1970 hronolog.html). Такая карта позволяет сделать выводы о влиянии Московского университета и его выпускников на жизнь России в различные периоды « от Ломоносова и до…» и может использоваться в образовательной деятельности.

Кластеры в сети текстуальных связей слов

Шайкевич Анатолий Янович
Институт русского языка им. В.В. Виноградова РАН
(Москва)

Корпус русской прозы (14 миллионов словоупотреблений) членится на фрагменты по 40 слов. Если совместная встречаемость двух слов во фрагментах существенно превышает величину, подсчитанную на основе нулевой гипотезы, делается вывод о наличии связи между этими словами. В ходе анализа получена огромная сеть текстуальных связей слов (более 30 тысяч слов и более 600 тысяч связей). В большой сети парных текстуальных связей неизбежно должны встречаться какие-то множества n слов. В той мере, в какой число связей приближается к максимальному n(n-1)/2, можно говорить о сгустке или **кластере** связей. Можно полагать, что сами кластеры связей обязаны своим существованием разным факторам — как общеязыковым, так и текстовым (в частности — сюжетным).

Едва ли увенчалась бы успехом попытка выработать какие-то строгие процедуры выделения кластеров. В докладе будут рассмотрены некоторые пути поиска кластеров, начиная от минимальных троек (вроде *безумный, бешеный, сумасшедший*, где число связей совпадает с максимальным), переходя далее ко все более крупным объединениям (*дуэль* с 58 словами, *доктор-болезнь* с более чем двумястами слов), кончая **мегакластерами** (*застолье, езда, комната* и т. п.).

Оглавление