

В. П. Захаров
*Санкт-Петербургский университет
Институт лингвистических исследований РАН
(Россия, Санкт-Петербург)
vz1311@yandex.ru*

КОРПУСА РУССКОГО ЯЗЫКА¹

Статья описывает предысторию российской корпусной лингвистики и корпуса русского языка, созданные как в России, так и за рубежом. Особое внимание уделяется Национальному корпусу русского языка, специализированным корпусам и диахроническому корпусу русских текстов Google books Ngram Viewer.

Ключевые слова: корпусная лингвистика, корпуса русского языка.

Введение

В последние годы создание корпусов и корпусно-ориентированные исследования стали неотъемлемой частью деятельности лингвистов. В мире корпусная лингвистика как особое направление сложилась к началу 1990-х годов. За прошедшие годы корпусная методология становится частью лингвистической науки, и все лингвисты, работающие в самых разных направлениях, как правило, проводят свои исследования на базе корпусов. Россия встала на этот «корпусный» путь с некоторым опозданием, но движется по нему очень быстро. Появляется большое число публикаций, посвященных созданию и использованию корпусов. Начинают выходить учебные пособия по корпусной лингвистике на русском языке (Захаров 2005, Гвишиани 2008, Захаров, Богданова 2011, 2013, Грудева 2012).

¹ Настоящая статья является расширенной и обновленной версией публикации на английском языке [Zakharov 2013].

Безусловно, ключевым моментом развития корпусной лингвистики в нашей стране стало создание Национального корпуса русского языка (2004) (<http://ruscorpora.ru>). О востребованности корпуса свидетельствуют многочисленные исследования и публикации, подготовленные на его основе. Часть из них описывается на сайте корпуса в разделе Studiorum (<http://studiorum.ruscorpora.ru/>). Однако корпуса русского языка сегодня создаются не только «в недрах» НКРЯ и не только в России. И если о НКРЯ сегодня написано много, то о других корпусах информации не хватает. С корпусной проблематикой во всем ее многообразии можно ознакомиться по материалам конференций «Корпусная лингвистика» (см. на сайте <http://corpora.phil.spbu.ru>), проводимой кафедрой математической лингвистики СПбГУ (Труды международной конференции «Корпусная лингвистика») и «Диалог» (<http://dialog-21.ru>) (Компьютерная лингвистика и интеллектуальные технологии: Материалы ежегодной Международной конференции «Диалог»). Однако был бы полезен единый очерк, дающий общее представление о корпусах русского языка. Есть, правда, обзоры корпусов славянских языков, включая русский [Резникова, Копотев 2005; Резникова 2009], то они не полны — за прошедшие годы корпусная лингвистика и «корпусостроение», естественно, ушли вперед.

Задача настоящего очерка — дать краткий общий обзор различных корпусов русского языка в России и за рубежом в их многообразии. За пределами данного обзора остаются детальное описание функциональных возможностей корпусных служб, вопросы использования корпусов для исследований по русскому языку, в целях обучения, настройки лингвопроцессоров и решения разнообразных лингвистических задач на базе корпусов.

1. Первые корпуса русского языка

Первый русскоязычный корпус был создан в 1980-е гг. в Университете Упсалы (Швеция). Однако еще до первых русскоязычных корпусов в полном смысле этого слова в 1960–70-е гг. был создан **Частотный словарь русского языка** под руководством Л.Н. Засориной [Засорина 1977], построенный на основе примитивных текстовых файлов, фактически, электронных словарных карточек, объемом в 1 млн словоупотреблений, включавших в себя лексику четырех жанров в примерно равной пропорции: общественно-политические

тексты, художественную литературу, научные и научно-популярные тексты из разных тематических областей и драматургию.

В процессе создания этих файлов — сегодня мы бы сказали, корпуса — решались все проблемы современной корпусной лингвистики, которые обсуждались и обсуждаются при создании полноценных корпусов:

- репрезентативность,
- сбалансированность,
- графематический анализ,
- нормализация,
- лемматизация.

Так что фактически это был первый корпус русского языка, не дошедший до наших дней.

В 1985 г. в СССР по инициативе академика А.П. Ершова (доклад 1978 г., см. [Ершов 1982]) были начаты работы по созданию **Машинного фонда русского языка** [Андрющенко 1989; Машинный фонд русского языка 1986]. Это был грандиозный проект. В создании фонда принимали участие более 40 организаций-соисполнителей, среди них Институт русского языка, Московский, Ленинградский (Санкт-Петербургский), Харьковский, Гродненский, Сыктывкарский и Саратовский университеты и др. В задачи фонда входило накопление на машинных носителях и в базах данных текстовых, лексикографических и грамматических источников, необходимых для научного изучения русского языка и для осуществления прикладных разработок. Одновременно создавались программные средства для проведения лингвистических исследований. В 1985–1992 гг. были осуществлены разработка концепции и архитектуры Машинного фонда русского языка, разработка концепции терминологического банка данных, введены в компьютер текстовые источники русской литературы XIX–XX вв., главные словари русского языка, краткая академическая грамматика, созданы текстовые корпуса поэзии, художественной прозы, общественно-политических и технических текстов. Однако с началом перестройки и в новых экономических условиях после 1991 г. работы по созданию фонда постепенно стали сокращаться и, наконец, совсем прекратились.

Уппсальский корпус русского языка (Uppsal'skij korpus russkix tekstov), созданный, как уже было сказано, в Университете Уппсалы, состоит из 600 текстов, его объем составляет 1 млн словоупотребле-

ний, поровну распределенных между образцами специальной и художественной литературы. По замыслу создателей, корпус должен был отражать современное состояние русского языка того периода. Цель формирования корпуса заключалась в том, чтобы представить, в первую очередь, письменный литературный язык.

В корпус отбирались тексты с 1985 по 1989 г. и художественные тексты с 1960 по 1988 г. Корпус составляли не фрагменты текстов, как в Брауновском корпусе, а целые тексты. В аннотации к корпусу отмечается, что среди специальных текстов особое внимание уделено более важным, с точки зрения создателей корпуса, темам, а среди художественных текстов предпочтение отдавалось более известным авторам. Тексты в корпусе записывались латиницей. Фрагмент корпуса выглядит следующим образом:

&Perestrojka vse glubhe zatragivaet hiznennye interesy millionov, obqestva v celom. Estestvenno, l~di xot,,t luŭwe u,,snit' sut' i naznaŭenie processov obnovleni,,, blihnje i dal'nie celi preobrazovanij, opredelit' svoe otnowenie k nim

Упсальский корпус сейчас входит в так называемые *«Тюбингенские корпуса русских текстов»*, созданные в рамках работы специального научно-исследовательского сектора SFB 441 Тюбингенского университета в 1990–2000-е гг. В целом тюбингенские корпуса представляют собой несбалансированное, морфологически размеченное (частично) собрание разнородных текстов, сегодня уже не соответствующее понятию лингвистического корпуса. Разметка была осуществлена при помощи статистического теггера (TnT). Поиск может производиться как по словоформам, так и — для размеченных текстов — по морфологическим тегам. Возможен вывод текста вместе с разметкой. Для ввода поискового выражения и вывода найденного текста можно выбрать одну из следующих кодировок: кириллицу (KOI8 или Windows-1251) или транслитерацию латинскими буквами. Поиск осуществляется при помощи программы CQP, представляющей собой систему для управления большими корпусами, разработанную Институтом машинной обработки языка Штутгартского университета.

Компьютерный корпус текстов русских газет конца XX в. (<http://www.philol.msu.ru/~lex/corpus/>) был создан на Филологическом факультете МГУ в 2000–2002 гг. в Лаборатории общей и

компьютерной лексикологии и лексикографии под руководством А.П. Поликарпова. Подбор обширного газетного материала для корпуса (тексты общим объемом более 11 млн словоупотреблений) был осуществлен на основе принципов включения в него полных номеров 13 российских газет на русском языке за отдельные даты 1994–1997 гг. (23110 текстов), представленности в нем газет ежедневных и неежедневных («МН», «Новая газета»), «левых» («Завтра», Правда», «Правда-5») и «правых», центральных и местных, общих и профессионально ориентированных (например, «Литературная газета»). Эти принципы позволяют получить относительно объективную и надежную картину соотношения в газетном материале текстов различного типа (например, различных жанров и жанровых типов), их единиц и отношений между ними.

Корпус создан, анализируется и управляется на основе системы Диктум-1, разработанной в Лаборатории общей и компьютерной лексикологии и лексикографии МГУ. С помощью этой системы тексты и единицы корпуса автоматически и полуавтоматически маркируются различного рода маркерами: тексты – маркерами газеты-источника, объема текста, его жанра, даты публикации и т. п.; словоупотребления – маркерами грамматических, лексических, морфемных и иных категорий. С подробным описанием корпуса и результатами частотно-статистического анализа газетных текстов можно ознакомиться на сайте (http://www.philol.msu.ru/~lex/corpus/corpus_descr.html).

Однако в открытом доступе исследователям доступна лишь незначительная часть корпуса. При подготовке демонстрационного варианта корпуса для Интернета был выделен фрагмент корпуса общим объемом более 200 тыс. словоупотреблений, проведена автоматическая лемматизация и морфологическая квалификация словоупотреблений корпуса, а также морфемная сегментация словоформ и лексем. В настоящее время корпус представляет лишь ограниченный интерес.

2. Современные корпуса русского языка

В начале этого века была осознана необходимость создания современного представительного корпуса русского языка. Как пилотные версии его можно рассматривать проекты корпусов Русский стандарт и БОКР (Большой Корпус русского языка) (<http://>

bokrcorpora.narod.ru/) (<http://corpus.leeds.ac.uk/serge/bokrcorpora/index.html>), который позиционировался как русский аналог Британского национального корпуса.

При этом следует упомянуть также большой корпус, созданный С. А. Шаровым в конце 1990 — начале 2000-х гг., недоступный в то время как корпус, но использованный как источник нового частотного словаря русского языка (<http://www.artint.ru/projects/frqlist.php>). Корпус включал в себя подборку прозы, политических мемуаров, газет и научно-популярной литературы того времени (около 40 миллионов слов, проза составляет примерно чуть больше половины объема). Все тексты корпуса были написаны на русском языке в промежутке между 1970 и 2002 гг.; большинство относится к периоду 1980–1995 гг., газетный корпус охватывал период 1997–1999 гг.

И, наконец, с апреля 2004 г. в открытом доступе появился **Национальный корпус русского языка** (НКРЯ) (<http://ruscorpora.ru/>), отвечающий критерию репрезентативности и всем другим требованиям, предъявляемым к современным корпусам. Объем корпуса составляет более 500 млн словоупотреблений (по данным на сентябрь 2014 г.). Жанровое разнообразие составляющих его текстов, которые относятся ко всем основным сферам использования русского языка (научной, официально-деловой, публицистической, церковно-богословской, художественной, разговорно-бытовой, включая устную и электронную коммуникацию) обеспечивает его сбалансированность. Разметка корпуса и подкорпусов, программное и документационное обеспечение позволяют решать сложные лингвистические задачи.

В состав корпуса входят основной корпус, в том числе подкорпус со снятой омонимией, газетный корпус, диалектный, синтаксический, обучающий, поэтический, устный, мультимедийный, исторический и параллельные корпуса. Структурно-статистическое наполнение корпуса см. <http://ruscorpora.ru/corpora-stat.html>.

НКРЯ хорошо известен и подробно описан в многочисленных публикациях (прежде всего [Национальный корпус русского языка 2005; Национальный корпус русского языка 2009], см. также <http://ruscorpora.ru/corpora-biblio.html>). Здесь в числе целого ряда его особенностей, отличающих его от других корпусов и в особенности национальных корпусов, отметим в первую очередь три: *мультиме-*

дийный подкорпус (<http://ruscorpora.ru/search-murco.html>), семантическая разметка (<http://ruscorpora.ru/corpora-sem.html>) и статистический инструмент проведения диахронических исследований «Графики» (<http://ruscorpora.ru/ngram.html>), подобный сервису Google Books Ngram Viewer.

Мультимедийный русский корпус (МУРКО) [Гришина 2005] образован фрагментами кинофильмов 1930–2000-х гг. и другими материалами, представленными в виде параллельных видеоряда, аудиоряда и текстовой расшифровки звучащей речи, а также специально размеченных наблюдаемых в кадре жестов. Возможен поиск не только по произносимому тексту, но и по жестам (кивание головой, похлопывание по плечу и т. п.) и типу речевого действия (согласие, ирония и т. п.). В поисковой выдаче видеофрагменты доступны для просмотра и прослушивания.

Семантическая разметка [Кусова и др. 2006] приписывает единицам текста один или несколько семантических и словообразовательных признаков, например, 'лицо', 'вещество', 'пространство', 'скорость', 'движение', 'обладание', 'свойство человека', 'диминутив', 'отглагольное имя' и т. п. Используется фасетная классификация, при которой одно слово может попадать в несколько классов. В основу семантической разметки положена система классификации русской лексики, принятая в базе данных «Лексикограф», которая разрабатывалась с 1992 г. в Отделе лингвистических исследований ВИНТИ РАН под рук. Е. В. Падучевой и Е. В. Рахилиной. И по указанным признакам, разбитым по нескольким категориям: разряд, таксономия, мереология, оценка, словообразование и т. п. — в корпусе осуществляется поиск. Можно искать контексты по основному значению слова или по вспомогательному. Можно найти все приставочные глаголы движения (отметить признак «движение» в категории «Таксономия», грамматический признак «глагол» и «тип морфемы: префикс» в категории «Словообразование») и т. д.

На сервисе «Графики» мы остановимся отдельно в разделе, посвященном диахроническим корпусам (п. 6.2).

Корпус русского литературного языка (1 млн словоупотреблений) (<http://www.narusco.ru/>) задумывался как интересный проект: тексты только второй половины XX в., сбалансированный жанровый состав, восстановленная буква «ё», проставленные словесные ударения. Однако практически корпус «не живет». Вначале разра-

ботчики не справились с задачей морфологической разметки, а затем просто перестали его развивать и поддерживать. И, как следствие, корпус доступен в сети, но не используется.

Открытый корпус (OpenCorpora) (<http://opencorpora.org/>) — это проект по созданию размеченного корпуса текстов силами лингвистического Интернет-сообщества. Корпус доступен бесплатно и в полном объеме (под лицензией CC-BY-SA). Это система, предназначенная для ввода текстов с лингвистической разметкой, предоставляющая интерфейс разметки, редактирования и исправления ошибок, инструмент для контроля качества и формат разметки для русского языка. Объем корпуса 1,3 млн. токенов, подкорпус со снятой омонимией — 18 тыс. токенов (сентябрь 2014 г.).

Работы начались в 2009 году, и сегодня можно констатировать, что первоначальные задачи: создать морфологически, синтаксически и семантически размеченный корпус текстов на русском языке, в полном объеме доступный для исследователей и редактируемый пользователями, — не решены. Тем не менее, корпус сыграл и играет положительную роль в развитии инструментов корпусной лингвистики. Кроме свободных ресурсов, первым среди которых следует назвать постоянно пополняемый морфологический словарь, нужно упомянуть, что корпус используется как учебная площадка студентами, обучающимися в области автоматизированной обработки текста.

Хельсинский аннотированный корпус (ХАНКО) (см. [Резникова, Копотев 2005]) создан в Хельсинском университете в начале 2000-х гг. как часть проекта «Функциональный синтаксис русского языка» (рук. проф. А. Мустайоки) и постоянно развивается. Объем корпуса — 100 тыс. словоупотреблений. Доступен в Интернете по адресу: <http://www.ling.helsinki.fi/projects/hanco/>. Корпус создан на основе статей первых четырех номеров журнала «Итоги» за 2001 г. В корпусе реализованы морфологическая и синтаксическая разметки и, соответственно, морфологический и синтаксический поиск. Особенность корпуса — тщательно проработанный формат лингвистического описания данных и полная визуальная (ручная) проверка результатов автоматической разметки, имеющая следствием полное снятие грамматической омонимии, там, где она может быть снята человеком. Синтаксическая разметка корпуса должна совмещать разметку в терминах членов предложения (уже реализована

см. набор синтаксических параметров на рис. 1) и в терминах деревьев зависимостей [Мустайоки и др. 2005].

Параметры предложений	Параметры клауз
<p style="text-align: center;"><u>Простое</u></p> <p><input type="checkbox"/> Простое предложение</p> <p style="text-align: center;"><u>Сложное</u></p> <p><input type="checkbox"/> Предложение с сочинительной связью</p> <p><input type="checkbox"/> Предложение с подчинительной связью</p> <p><input type="checkbox"/> Предложение с бессоюзной связью</p>	<p style="text-align: center;"><u>Роль</u></p> <p><input type="checkbox"/> Самостоятельная</p> <p><input type="checkbox"/> Главная</p> <p><input type="checkbox"/> Подчиненная</p> <p style="text-align: center;"><u>Структура</u></p> <p><input type="checkbox"/> Двусоставная</p> <p><input type="checkbox"/> Односоставная</p> <p><input type="checkbox"/> Фразеологизированная</p> <p><input type="checkbox"/> Эллиптическая</p> <p><input type="checkbox"/> Эллиптическая</p>
Выбор синтаксических параметров	Параметры клаузы для слова 1
<p style="text-align: center;"><u>Части именного сказуемого</u></p> <p><input type="checkbox"/> Связочная часть</p> <p><input type="checkbox"/> Присвязочная часть</p> <p style="text-align: center;"><u>Главные члены предложения</u></p> <p><input type="checkbox"/> Главный член односоставного предложения</p> <p><input type="checkbox"/> Подлежащее (= подлежащее двусоставного предложения)</p> <p><input type="checkbox"/> Сказуемое (= сказуемое двусоставного предложения)</p> <p style="text-align: center;"><u>Второстепенные члены предложения</u></p> <p><input type="checkbox"/> Дополнение</p> <p><input type="checkbox"/> Определение</p> <p><input type="checkbox"/> обстоятельство</p> <p style="text-align: center;"><u>Не является членом предложения</u></p> <p style="text-align: center;">Обращение</p>	<p style="text-align: center;"><u>Роль</u></p> <p><input type="checkbox"/> Самостоятельная</p> <p><input type="checkbox"/> Главная</p> <p><input type="checkbox"/> Подчиненная</p> <p style="text-align: center;"><u>Структура</u></p> <p><input type="checkbox"/> Двусоставная</p> <p><input type="checkbox"/> Односоставная</p> <p><input type="checkbox"/> Фразеологизированная</p> <p><input type="checkbox"/> Эллиптическая</p> <p><input type="checkbox"/> Эллиптическая</p>

Рис. 1. Набор синтаксических параметров в корпусе ХАНКО.

Кроме того, размечены многословные устойчивые обороты (примерно 2000 единиц). Планируется семантическая разметка в терминах семантических категорий. Цели и особенности корпуса описаны на веб-странице проекта <http://www.helsinki.fi/venaja/russian/e-material/hanco/index.htm>. Корпус преследует в первую очередь учебные цели.

Корпуса университета г. Лидс (Великобритания). В течение 2000-х гг. в университете г. Лидс в Центре переводческих исследований С.А. Шаровым создано большое количество корпусов для разных языков (английский, арабский, китайский, французский, немецкий, итальянский, японский, испанский, польский и др.) (<http://corpus.leeds.ac.uk/>). Среди них имеются и корпуса русского языка (<http://corpus.leeds.ac.uk/ruscorpora.html>). Это, в первую очередь, две версии Национального корпуса русского языка объемом в 116 млн словоупотреблений (с разной морфологической разметкой). На основе этого подмножества НКРЯ был создан Частотный словарь русского языка [Ляшевская, Шаров 2009]. По объему эта версия уступает НКРЯ в его сегодняшнем состоянии, но интересна функциональными возможностями корпусной системы, в среде которой функционирует этот русскоязычный массив. Кроме того, на сайте Университета Лидса представлены и другие русскоязычные корпуса: корпус русских газет (2001–2004 гг., 77 млн словоупотреблений), корпус русских текстов из Интернета (198 млн словоупотреблений), бизнес-корпус (15 млн словоупотреблений), корпус на основе русской Википедии (178 млн токенов), устный корпус (русскоязычные веб-форумы, 1 млн), корпус научных текстов (5,1 млн), корпус из текстов «Живого журнала» (700 млн) и большой 2-миллиардный русскоязычный корпус, созданный на базе Интернета по технологии wasky (эта проблематика — новая технология создания корпусов — находится за пределами данной статьи; см. материалы семинаров <https://sigwac.org.uk/>).

Поисковый интерфейс Leeds CQP базируется на корпусном менеджере IMS Corpus Workbench и предоставляет интересные возможности (рис. 2).

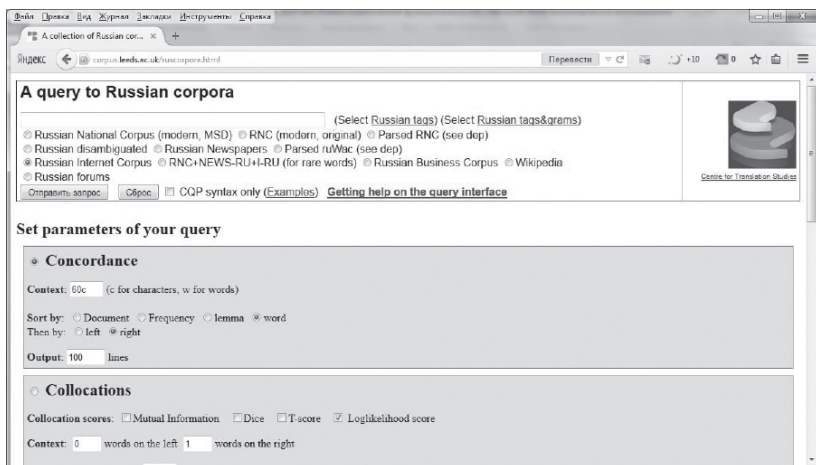


Рис. 2. Интерфейс доступа к русским корпусам в Университете Лидса.

Он позволяет вести точный лексико-грамматический поиск на основе специального языка запросов, в том числе с использованием языка регулярных выражений. Имеются способы управления выходным интерфейсом, формой представления результатов поиска. Можно также получить списки коллокаций, вычисленных и упорядоченных на основе ассоциативных мер MI, Dice, T-score и Log-likelihood. Там же имеется коллекция различных программных средств для обработки корпусных текстовых данных (<http://corpus.leeds.ac.uk/tools/>).

Новая система IntelliText (<http://corpus.leeds.ac.uk/it>) имеет более развитый и дружелюбный интерфейс (рис. 3, 4) и расширяет набор функций, в частности, предоставляется возможность создавать свои корпуса.

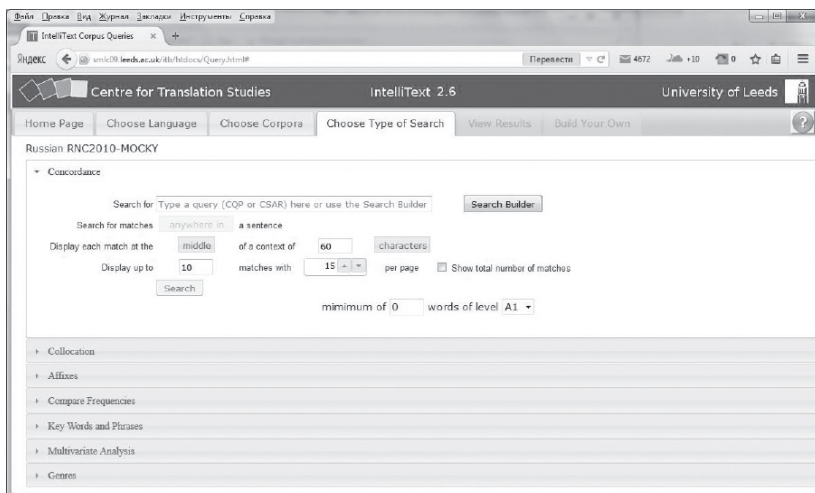


Рис. 3. Интерфейс поиска системы IntelliText: конкорданс

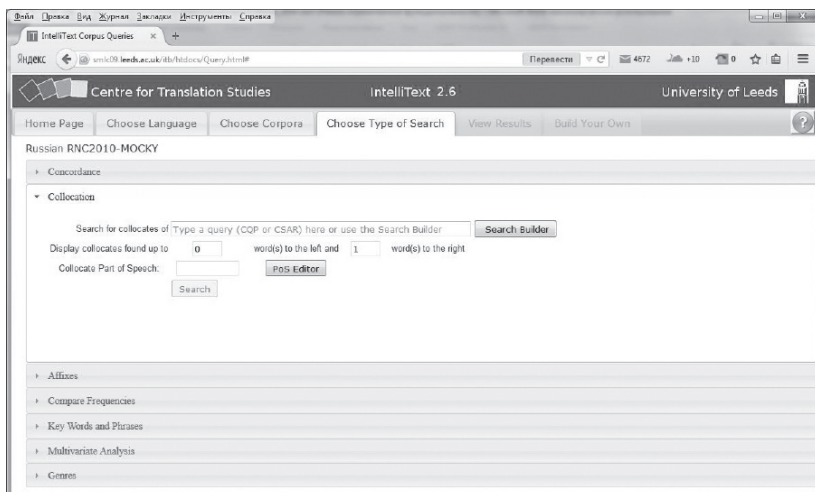


Рис. 4. Интерфейс поиска системы IntelliText: коллокации.

Корпус Библиотеки Мошкова. На сайте группы АОТ (<http://aot.ru/search1.html>) имеется большой корпус русских текстов объемом 680 млн словоупотреблений (844 млн токенов), созданный А. В. Сокирко по текстам из библиотеки Мошкова. Можно искать по лексическим единицам с учетом частей речи и морфологических характеристик, используя мощный язык запросов корпусного менеджера DDC (<http://aot.ru/concor.html>). Там же на сайте АОТ имеется сервис поиска биграмм (54 млн), вычисленных по мере MI, созданный А.В. Сокирко и А.Н. Авериним по текстам корпуса Библиотеки Мошкова (<http://aot.ru/demo/bigrams.html>).

Корпуса в системе Sketch Engine. Английская лингвистическая служба Lexical Computing Ltd. (A. Kilgarriff) предоставляет на коммерческой основе доступ к 329 корпусам (сентябрь 2015 г.) различных языков мира. Среди них имеется ряд корпусов русского языка и в их числе большой корпус, созданный из текстов Интернета по технологии wasky объемом 18 млрд токенов (14,5 млрд текстоформ) (ruTenTen 2011) (рис. 5, 6). Английские исследователи совместно с чешскими разработчиками из Университета им. Масарика (Брно) создали многофункциональную корпусную систему Sketch Engine (<http://sketchengine.co.uk/>) [Kilgarriff 2004], обладающую многими уникальными возможностями. Помимо стандартного поиска с выдачей конкорданса она выдает списки коллокаций, сформированные на основе 7 мер ассоциации, списки коллокаций по отдельным синтаксическим моделям (word sketches), формирует частотные списки слов, словоформ и тегов, группирует лексические единицы в лексико-семантические поля с внутренней кластеризацией и указанием силы связи между лексемами.

Corpora

- Create corpus
- WebBootCat
- Upload TMX

Parallel corpora

- Compare corpora
- Configuration templates
- Sketch grammars
- Subcorpus definitions
- User groups

Admin

Local administration

Support

Help index

Report an error

Request a feature

Corpora

Language	Corpus name	Tokens	Words
Estonian	EstonianNC	563,220,548	463,827,780
Estonian	etTenTen	330,045,196	260,559,829
French	frTenTen12	12,369,868,562	10,666,617,369
Russian	ruTenTen11	20,162,118,568	15,763,181,803
Russian	ruTenTen11 (RFTagger, sample 50M) with term grammar	59,115,079	45,955,925
Russian	ruTenTen11 (sample 50M)	59,211,785	46,040,989
Russian	ruTenTen11 (sample 50M) with term grammar	59,211,785	46,040,989
Russian	ruTenTen11 (2011, RFTagger, filtered)	18,280,485,876	14,553,856,113

Featured corpora | All corpora | Parallel corpora

My corpora

Language	Corpus name	Configuration template	Tokens
English	EnglishLinguistics	TreeTagger for English	34,584
English	Test April English2	TreeTagger for English	8
English	Twilight	TreeTagger for English	57,261
Estonian	Estonian_web	TreeTagger for Estonian	75,390
Estonian	Jelena2	General tokeniser	9,469
Russian	Bees	RFTagger/TreeTagger for Russian	0
Russian	bokrfonok	bokrfonok	19,561,948
Russian	Corpus Linguistics	TreeTagger for Russian	343,444
Russian	CL2004-2008	TreeTagger for Russian	225,574
Russian	CL2004-2011	RFTagger/TreeTagger for Russian	273,270
Russian	Dialog	TreeTagger for Russian	2,611,111
Russian	Russian Web Corpus	General tokeniser	9,493,299
Russian	Homoeopathy_test	TreeTagger for Russian	2,102,352

Рис. 5. Интерфейс доступа к корпусам Sketch Engine.

Concordance

- Word List
- Word Sketch
- Thesaurus
- Find X
- Sketch-Diff
- Corpus Info
- ?

Query types: Context Text types

Query type: ☒ simple ☐ lemma ☐ phrase ☐ word ☐ character ☐ CQL

Lemmas: PoS: unspecified

Phrases:

Word Forms: PoS: unspecified ☐ match case

Character:

CQL: Default attributes: word Target summary

Lexical Computing
 Sketch Engine (ver:2.28.3-SAE-2.109.9-3.48.15)
 Interface language: English | Český | 简体中文 | 繁體中文 | Español | slovenščina | hrvatski

Рис. 6. Интерфейс системы Sketch Engine: конкорданс.

Корпуса в системе Aranea (A Family of Comparable Gigaword Web Corpora). Это система псевдопараллельных корпусов для 14 европейских языков, включая русский (15-м языком выступает китайский), созданная в Университете им. А. Коменского в Братиславе (<http://ucts.uniba.sk/>). Все языки представлены корпусами двух ти-

пов, созданными по технологии wacky: Maius (1200 млн токенов; как правило, из этого числа около 1000 млн токенов представляют собой слова, или в другой терминологии, текстоформы) и Minus (120 млн токенов, чуть более 90 млн текстоформ). Для некоторых языков имеются региональные варианты, так, англоязычных корпусов насчитывается 6: Araneum Anglicum, Araneum Anglicum Asiaticum, Araneum Anglicum Africanum, каждый соответственно, Maius и Minus. То же самое относится и к русскоязычным корпусам, которых тоже 6: Araneum Russicum Maius & Minus (русский универсальный), Araneum Russicum Russicum Maius & Minus (русский на основе текстов с сайтов с доменом .ru), Araneum Russicum Externum Maius & Minus (русский на основе текстов с доменами, отличными от домена .ru). Подробнее см. [Benko 2013]. Корпуса поддерживаются системой NoSketch Engine, отличающейся от Sketch Engine уменьшенной функциональностью, но зато бесплатно распространяемой (<http://nlp.fi.muni.cz/trac/noske>).

3. Лингвистические ресурсы русской устной речи

Выделяется 2 основных типа корпусов: 1) корпуса текстов (расшифровок) устной речи, которые аннотируются по правилам, схожим с разметкой текстовых корпусов, 2) мультимедийные (звуковые, речевые) корпуса или корпуса звучащей речи, единицей описания которых являются собственно звукозаписи устной речи, которые расшифровываются (транскрибируются) и далее аннотируются в соответствии с поставленными перед исследователями задачами.

Наиболее представительным ресурсом первого типа является **Корпус устных текстов НКРЯ** объемом более 10 млн словоупотреблений, хорошо известный по публикациям (см., например [Гришина 2005; Гришина, Савчук 2009]), поэтому здесь мы его не рассматриваем.

Корпуса звучащей русской речи, или звуковые корпуса активно создаются последние 50 лет в научно-исследовательских центрах и лабораториях для поддержки задач речевых технологий (синтеза и распознавания речи) (например, разработки НИИ «Дальняя связь», Института физиологии им. Павлова в советские годы, в наше время — Центра речевых технологий и др.)

Известны записи звучащей речи, ориентированные на проведение фундаментальных научных исследований, например, Фонети-

ческий Фонд русского языка, разрабатываемый на кафедре фонетики СПбГУ (ЛГУ) в конце прошлого века [Бондарко 1988] и многочисленные лингвистические ресурсы, содержащие аудиоматериалы по языкам малых народов Российской Федерации. Большинство таких ресурсов ориентированы на запись ограниченного набора фраз (текстов, слов), осуществляемых в лабораторных условиях или дистантно (например, по телефону).

Из разработок последних лет выделяются два ресурса — корпус звучащей речи «Рассказы о сновидениях» и корпус повседневной русской речи «Один речевой день».

«Рассказы о сновидениях» представляет собой небольшой (около 2 часов звучания) и жанрово ограниченный (рассказы детей о своих сновидениях), но хорошо проаннотированный и проанализированный ресурс, ориентированный на исследование русской устной речи и дискурса [Кибрик, Подлесская 2009]. В последние годы ресурс расширяется за счет привлечения других звуковых материалов (<http://spokencorpora.ru>).

Наиболее представительным корпусом повседневной устной русской речи на настоящий момент является речевой корпус **Один речевой день** (ОРД), разрабатываемый на Филологическом факультете СПбГУ [Asinovsky et al. 2009]. Корпус создается с целью изучения реальной речи носителей языка в естественных условиях коммуникации, и в этом его отличие от абсолютного большинства речевых корпусов, записанных в лабораторных и других специальных условиях.

Первая серия звукозаписей осуществлена осенью 2007 г. Для этого была отобрана группа информантов из 30 человек, представляющих разные социальные и возрастные слои населения Санкт-Петербурга и давших согласие прожить один день с «диктофоном на шее». Информанты получили подробный инструктаж о методике проведения звукозаписи своих речевых контактов в течение суток, заполнили социологические анкеты и прошли психологическое тестирование. Помимо речи информантов, в корпусе представлены записи их коммуникантов (родственников, друзей, коллег, знакомых и незнакомых), среди которых были люди самого разного возраста и разных специальностей. Общая длительность записанного материала — более 500 часов. Расшифровке и многоуровневой разметке подвергнуто 50 часов (по данным на сентябрь 2014 г.). Сплошное

аннотирование проводится, в частности, по следующим базовым уровням: 1) реплики (фразы), 2) говорящий, 3) невербальные речевые события, 4) фонетический комментарий, 5) фразовый комментарий, 6) мини-эпизод речевой коммуникации и др. [Шерстинова и др. 2009].

Корпус постоянно расширяется, проводятся новые звукозаписи, осуществляется их расшифровка и увеличивается количество уровней аннотирования. На конец 2016 года планируется увеличение объема текстовых расшифровок до 1 млн словоупотреблений. Поиск по анонимизированным расшифровкам и материалам корпуса ОРД будет доступен на сайте <http://www.ord-corpus.spbu.ru>, начиная с 2016 г., публикация самих звукозаписей не планируется ввиду большого объема личной и частной информации, представленной в данном ресурсе.

Данный звуковой корпус позволяет изучать русскую повседневную речь на всех лингвистических уровнях. Например, можно исследовать лингвистическую динамику записанного материала: изучать временные ряды количественных переменных с помощью стандартных статистических методов и анализировать частотные ряды (лексики, грамматических и, в частности, синтаксических структур, семантики или разговорных тем, тех или иных акустических явлений или просодических контуров) в зависимости от времени суток и условий коммуникации в самом широком понимании этого термина. Можно решать множество других задач, таких как анализ влияния профессии на бытовую жизнь человека, получение информации о среднем артикуляционном темпе спонтанной речи носителей русского языка.

Результаты исследований, проведенных на материале корпуса ОРД, опубликованы в многочисленных публикациях [см., напр., Асиновский и др. 2010].

Мультимодальные корпуса включают видеозапись участников коммуникации, поэтому с их помощью можно исследовать эмоции.

Мультимедийный русский корпус в составе НКРЯ мы уже упоминали, и с ним можно познакомиться по публикациям [Гришина 2005; 2009], современное состояние и перспективы развития описаны в статье Е.А. Гришиной в настоящем сборнике.

Русскоязычный эмоциональный корпус (REC), размеченный с учетом данных о мимике, движениях рук, бровей и т. д., позволя-

ет изучить стратегии эмоционального взаимодействия и конфликта, непрерывное коммуникативное поведение, гезитации и речевые сбои и др. (<http://www.harpia.ru/rec>) [Котов, Гопкало 2011]. Он может также использоваться как материал для обучения работников клиентских служб или как база данных эмоциональных реакций для мультипликаторов и режиссеров.

В Иркутском государственном лингвистическом университете идет работа по созданию *Учебного Мультимодального Корпуса* (УМКО), включающего видеозаписи неподготовленных учебных диалогов носителей и «неносителей» русского и китайского языков по определенным темам, размеченных в программе ELAN и представленных также в виде параллельных корпусов, выровненных по смысловым блокам внутри диалогов [Богданова 2013]. Например, диалог носителей русского языка на русском языке сопоставляется с диалогом на ту же тему на русском языке («Знакомство», «Регистрация в аэропорту» и др.) китайцев, изучающих русский язык. Данный корпус предназначен, в первую очередь, для лингводидактических целей, так как позволяет выявить типичные ошибки и найти пути их устранения в ходе учебных занятий и самостоятельной работы студентов.

Назовем в этом разделе еще *Корпус устных рассказов* на русском языке [Николаева 2009], стимулом для которых послужил 6-минутный видеосюжет, так называемый «Фильм о грушах» (“Pear film”). Об этом фильме было записано 8 рассказов, сделанных студентами МГУ. Общая длительность звучания — около 2 часов; объем корпуса — около 14 тысяч словоупотреблений. Всего в корпусе было 595 элементарных дискурсивных единиц, которые обычно совпадают с простым предложением, и 327 иллюстративных жестов, которые, в соответствии с подходом Г.Е. Крейдлина, понимаются как носители информации, выступая в качестве знаковых кинетических единиц выражения и передачи информации. На примере из корпуса устных рассказов исследователям удалось показать, как отдельные признаки жестов и положения рук могут добавлять дополнительную информацию касательно организации дискурса, состояния говорящего и процесса коммуникации. Так, изменение положения покоя рук между жестами достаточно последовательно указывает на границу между сегментами нарратива. Данный пример демонстрирует предоставляемые мультимедийным корпусом возможности изучения связи структуры устного нарратива и иллюстративных жестов.

4. Специальные корпуса текстов

Специальные корпуса текстов — это сбалансированные корпуса, как правило, небольшие по размеру, подчиненные определенной исследовательской задаче и предназначенные для использования преимущественно в целях, соответствующих замыслу составителя. Второе понимание специальных корпусов — это тексты по определенной тематике, относящиеся к какой-то предметной области, или тексты определенного типа (например, патенты).

Примером специального корпуса текстов может быть *Санкт-Петербургский учебный корпус текстов школьников, изучающих английский язык* (SPbEFLLC), созданный на кафедре прикладной лингвистики РГПУ им. А.И. Герцена [Камшилова 2012]. Основной целью его создания было исследование особенностей английских текстов, порождаемых русскими школьниками. Аутентичный текстовый материал был собран в школах Санкт-Петербурга в период с ноября по декабрь 2007 г. Авторами текстов являются 78 учеников 9–11 классов, предварительно прошедших тестирование. Уровень владения английским языком был определен как средний/intermediate (26%) и выше среднего/upper-intermediate (74%). Размер данного корпуса составляет около 50 тыс. словоупотреблений.

Исследование на базе корпуса показало, что систематическое предпочтение максимально простых структур развернутым и более естественным моделям стандартного английского языка приводит к так называемой «структурной бедности» речевых произведений носителей языка. В репертуаре грамматических структур, обнаруженных в SPbEFLLC, есть такие, которые представляют собой случаи «переходной грамматики» (интеръязыка), выражающиеся, например, в нарушении правил наполнения компонентов базовых структур. Так формируется ядро грамматики EFL (English as a Foreign Language), которое не совпадает с базовыми грамматическими структурами литературного английского языка. На основании корпусных данных авторы высказывают предположение о том, что складывающиеся нормы «глобального английского» во многом опираются на «окаменевшие» модели интеръязыка.

Сложным объектом с точки зрения создания и стандартизации являются исторические корпуса, такие как, например, *Санкт-Петербургский корпус агиографических текстов* (СКАТ) [Герд и др. 2004], доступный на сайте <http://project.phil.pu.ru/skat>. СКАТ —

это электронный корпус текстов по памятникам древнерусской агиографической литературы XV–XVII вв., созданный на кафедре математической лингвистики филологического факультета СПбГУ. Язык агиографических произведений во многом обусловил судьбу и характер русского литературного языка XV–XVII вв. Отображение этого языка является первостепенной задачей создаваемого корпуса текстов русских житий того времени, что достигается, в частности, за счет широкого географического охвата территорий, где в разное время создавались памятники русской агиографии. К настоящему времени корпус охватывает 60 житий, их общий объем — более 500 тыс. словоупотреблений.

Корпус является базой широкомасштабного проекта по изданию уникальной серии текстов «Памятники русской агиографической литературы». Для представления рукописей в корпусе была разработана система отображения древнерусской графики, которая позволяет воспроизводить текст с высокой степенью приближения к оригиналу. Отображаются графические начертания всех древнерусских букв и их семантически значимых вариантов. Разработана специальная программа, позволяющая получать к введенным текстам (к каждому в отдельности или к нескольким вместе) указатели словоформ, то есть списки словоформ с их адресами (номерами листов и строк) в рукописях, фактически, конкорданс.

В настоящее время осуществляется грамматическая разметка представленных в корпусе житий [Алексеев и др. 2011]. В формальном плане разметка корпуса основывается на международных нормах оформления электронных изданий текста, в частности, на рекомендациях проекта Text Encoding Initiative (TEI), однако содержит также и дополнительные элементы, которые необходимы для адекватного отображения особенностей русского рукописного текста.

Разработан формат представления грамматической информации в структуре XML. Для каждой словоформы текста указывается ее частеречная принадлежность и приводятся значения всех релевантных грамматических категорий. При этом учитываются некоторые особенности морфологии текстов XV–XVI вв. Дело в том, что тексты житий написаны на церковнославянском языке и, с одной стороны, сохраняют архаичные формы, уже вышедшие или выходящие из употребления в древнерусском языке того периода, а с другой

стороны, отражают живые языковые процессы, такие как смешение типов склонений существительных, формирование категории одушевленности, утрата двойственного числа, перестройка системы прошедших времен глагола, образование деепричастия и т. д. Поэтому в формате грамматической разметки предусмотрена возможность отражения переходных явлений: через косую черту приводятся ожидаемое значение соответствующей категории (тип склонения, падеж и т. п.) и реально встретившееся в тексте. Представленную таким образом информацию можно подвергать разным видам автоматического анализа, например, можно оценить степень архаичности или новизны того или иного текста.

Формат синтаксической разметки текстов житий ориентируется на систему аннотирования, принятую в НКРЯ в синтаксически размеченном корпусе СинТагРус [Алексеева 2014]. Очевидно, что синтаксис церковнославянского языка, на котором писались жития в Древней Руси, отличается от современного синтаксиса. Поэтому был составлен свой набор синтаксических отношений, в который вошли 39 отношений, используемых в СинТагРус, и 25 отношений, выделенных на основании анализа конкретных текстов житий и научной литературы. Работа над форматом продолжается.

В настоящее время на сайте корпуса в открытом доступе представлена лишь часть корпуса, а именно, в формате PDF представлено 13 житий, 10 из них представлены также и в формате XML.

Специальные корпуса текстов могут быть востребованы не менее, чем национальные. Любой *отраслевой специальный корпус текстов* может пригодиться и в данной конкретной отрасли (например, кораблестроение, металлы, экология, навигация и т. д.), и в смежных областях, поскольку он дает специалисту самое главное — термины в их профессиональном конкретном окружении, позволяет отследить изменения в терминологии, включая появление новых терминов. В СПбГУ ведется работа по созданию семейства представительных специальных корпусов общим объемом более 20 млн. словоупотреблений, охватывающих широкую тематику (компьютерная лингвистика, гомеопатия, радиотехника, шахматы, футбол, пчеловодство, растения, научно-популярные (журнал «Наука и жизнь»), путешествия) [Митрофанова и др. 2014].

5. Параллельные корпуса

По критерию «параллельность» корпуса делятся на одноязычные, двуязычные и многоязычные. В одноязычных корпусах противопоставляются диалекты, варианты языка. Двуязычные и многоязычные корпуса можно разделить на два основных типа:

1) корпуса, представляющие множество текстов-оригиналов, написанных на каком-либо исходном языке, и текстов-переводов этих исходных текстов на один или несколько других языков;

2) корпуса, объединяющие тексты из одной и той же тематической области, независимо написанные на двух или нескольких языках; такие корпуса помогают в работе с терминологией и часто используются переводчиками.

Корпуса обоих типов используются в целях разработки эффективных методов перевода, в том числе, машинного, а также для сравнительных исследований языков (в области лексикологии, грамматики, стилистики, переводоведения и т.д.). Естественно, среди параллельных корпусов есть корпуса, где в качестве одного из языков представлен русский. Не касаясь проблем создания таких корпусов (подбор материала, выравнивание и т.п.), просто перечислим известные параллельные корпуса, где есть тексты на русском языке.

Прежде всего, это **параллельные корпуса НКРЯ**, где на сентябрь 2014 г. для русского языка имеются следующие «параллельные» языки: английский, немецкий, французский, испанский, итальянский, польский, украинский, белорусский, армянский, болгарский и латышский языки. Подробнее см. [Добровольский и др. 2005]. При поиске используется стандартный интерфейс поиска в основном корпусе.

Корпус PARASOL (PARAllel corpus of Slavic and Other Languages) — параллельный корпус современных текстов, созданный в Университете Регенсбурга [Waldenfels 2006, 2011]. По состоянию на март 2014 г. корпус насчитывал более 27 млн токенов на 31 языке. Корпус размечен морфологически для большинства языков и доступен онлайн (<http://slavist.de/Ursynow/>).

Корпус InterCorp — часть Чешского национального корпуса (www.korpus.cz/intercorp) [Čermák, Rosen 2012]. В корпусе содержатся тексты на 38 языках, включая русский, образующие «параллель» с текстами на чешском языке. Объем русской части — 13,4 млн токенов. Все тексты выровнены и морфологически размечены. Для

поиска используется единый поисковый интерфейс Чешского национального корпуса. Корпус доступен онлайн для зарегистрированных пользователей.

Корпус PARUS (PAralelní RUsko-Slovenský korpus) — параллельный корпус современных текстов, созданный в Институте языковедения им. Л. Штура Словацкой академии наук [Гарабик, Захаров 2006]. Корпус состоит из переводов с русского на словацкий и со словацкого на русский. Тексты выровнены по предложениям. Все тексты корпуса морфологически размечены. Версия корпуса 2.0 (январь 2014 г.) насчитывает 8,45 млн токенов, 4,2 млн токенов в словацкой части и 4,25 млн токенов в русской. Корпус доступен онлайн по адресу <http://korpus.sk:8099/manatee.ks/index>.

Имеются также большой **польско-русский** и **русско-польский параллельный корпус** объемом 30 млн токенов, разработанный в Варшавском университете (<http://pol-ros.polon.uw.edu.pl/>), небольшой **болгарско-русский параллельный корпус** (211 текстов, 3,3 млн токенов), разработанный в рамках работ по созданию Болгарского национального корпуса в Институте болгарского языка Болгарской академии наук (http://www.ibl.bas.bg/BGNC_parallel_bg.htm; <http://search.dcl.bas.bg/bg/>).

Корпус ASPAC — многоязычный корпус художественных текстов (около 70 произведений) и их переводов (иногда до 6 переводов одного и того же текста) создан в Амстердамском университете (ASPAC — Amsterdam Slavic Parallel Aligned Corpus) (<http://www.uva.nl/over-de-uva/organisatie/medewerkers/content/b/a/a.a.barentsen/a.a.barentsen.html>). Несмотря на название, корпус включает тексты и на 11 неславянских языках. Тексты выровнены по абзацам. Корпус доступен в режиме загрузки по соглашению.

Параллельных корпусов достаточно много, но чаще всего они создаются для решения различных конкретных задач (переводоведение (см., например, [Добровольский 2003а, 2003б; Михайлов 2003]), литературоведение (см. Параллельный корпус переводов «Слова о полку Игореве», <http://www.nevmenandr.net/slovo/pro.php>), двуязычная и многоязычная терминология и др.) и не всегда доступны. Особенно много таких «целевых» корпусов создается сейчас в Европейском Союзе в рамках работ по терминологии (их обзор см. в [Steinberger R. et al. 2014]).

6. Диахронические корпуса

Выявление, описание и интерпретация изменений языка во времени — задача диахронического исследования. Еще недавно проведение такого исследования требовало больших усилий и затрат времени. Сегодня компьютерные технологии и корпусная лингвистика дают для него принципиально новые инструменты. Для этого создаются диахронические корпуса. Диахронические корпуса – понятие неопределенное. Можно сказать, что это корпуса, которые позволяют изучать развитие языка на протяжении какого-то (достаточно длинного) промежутка времени или изучать язык в его прежних состояниях. Таким образом, это понятие подразумевает представительный исторический корпус и (желательно) инструмент для диахронических исследований. В числе диахронических корпусов можно назвать уже упомянутый СКАТ, Регенсбургский диахронический корпус русского языка (древнерусские тексты) (<http://rhssl1.uni-regensburg.de/SlavKo/korpus/trudi-new>), Коллекцию древнейших и средневековых славянских и русских текстов «Манускрипт» (<http://mns.udsu.ru/>), Рукописные памятники Древней Руси: берестяные грамоты, летописи, рукописная книга (<http://gramoty.ru/>) и др.

В данной статье мы не рассматриваем особенности и наполнение разных диахронических корпусов, а остановимся лишь на программных инструментах, поддерживающих диахронические исследования, базирующиеся на корпусах. С точки зрения функциональности большой набор лингвистических модулей предлагает система «Манускрипт». Она позволяет знакомиться с текстами, указателями и осуществлять выборку данных, предлагает пользователям специализированный редактор для ввода, редактирования и фрагментирования текстов, модуль выборок и запросов, позволяющий подготовить данные для лингвистических, палеографических и текстологических исследований, морфологический анализатор для автоматического анализа и синтеза словоформ древнерусского языка и др. Здесь мы опишем два инструмента, находящихся непосредственно в рамках корпусной лингвистики и выполняющих статистическую обработку корпусных данных.

6.1. Сервис Google Books Ngram Viewer

Первой из таких специальных диахронических систем был сервис Google Books Ngram Viewer (<http://books.google.com/ngrams>) [Michell

2011; Захаров, Масевич 2014]. Эта система является в настоящее время наиболее мощным инструментом диахронических исследований. Доступ к ней открыт, начиная с 2009 года. Google books Ngram Viewer представляет собой информационную систему, которая содержит несколько корпусов размеченных текстов книг на 9 языках. Система также содержит отдельно корпус британского и американского английского языка, корпус всех вариантов английского языка. Самые поздние тексты корпуса, доступные для пользователей в настоящее время, относятся к 2008 году. На конец 2013 г. база данных насчитывала более 8 млн книг (текстов), что составляет около 6% всех когда-либо опубликованных печатных книг. Корпус книг на русском языке содержит 591 310 книг (текстов) или 67 137 666 353 словоупотреблений.

Интерфейс системы. Для каждой заданной лексической единицы для заданного временного интервала строится график. Каждая кривая графика маркируется цветом, в конце кривой указывается, какой N-грамме (слову или словосочетанию) она соответствует (рис. 7). Под термином N-грамма понимается последовательность от одного до пяти слов, причем N-грамма должна встречаться в корпусе не менее 40 раз. По вертикальной оси графика откладывается относительная частота встречаемости заданной N-граммы в данном году, выраженная в процентах. На горизонтальной оси показаны годы, входящие в заданный временной интервал.

Google books Ngram Viewer

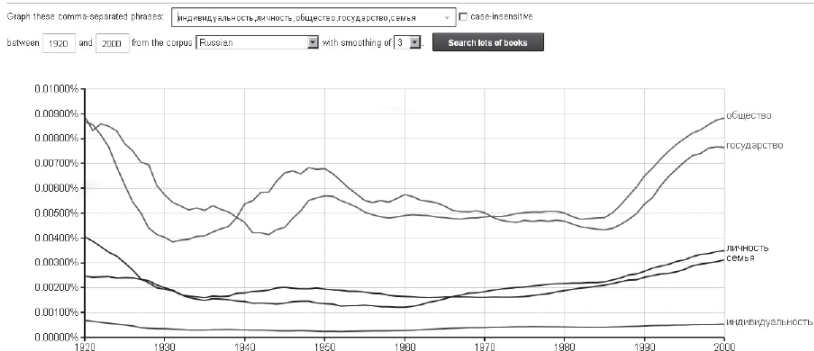


Рис. 7. Общий вид экрана с результатами выполнения запроса.

При построении графиков, показывающих динамику изменения частоты употребления, используется так называемое «сглаживание» (smoothing). При нулевом сглаживании в графике учитывается относительная частота встречаемости N-граммы за каждый год. Тенденция в динамике встречаемости слов прослеживается более отчетливо при скользящем усреднении данных. В окне “Smoothing” интерфейса, по умолчанию указано сглаживание 3, что означает, что для данного года для данной лексической единицы высчитывается среднее значение из семи чисел — число употреблений слова в данном году и число употреблений для трех предыдущих и трех последующих лет.

Возможно определение координат любой точки графика. Для этого достаточно установить курсор на любую точку над нужным годом. Система в этом случае выдаст сообщение о вертикальной и горизонтальной координатах этой точки для всех кривых. Если же установить курсор непосредственно на кривую, то исследуемая кривая будет выделена (рис. 8).

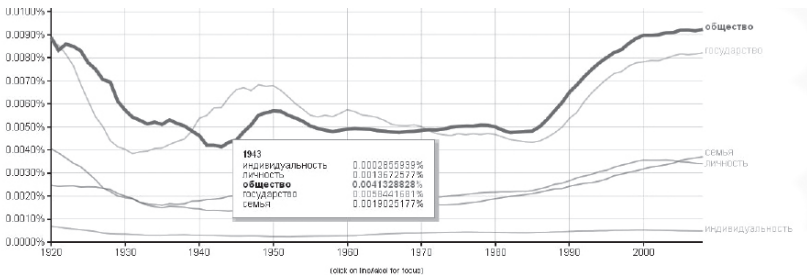


Рис. 8. Сообщение о координатах кривых для выбранного года (label for focus).

Кроме построения графиков, система представляет ссылки к текстам, найденным по запросам. Как правило, это библиографические описания книг и фрагменты текстов с выделением в них заданных N-грамм. В некоторых случаях доступен полный текст книги в графическом формате.

Поиск и лингвистические особенности системы

Для задания запроса на построение графика заполняется специальная форма (рис. 9).

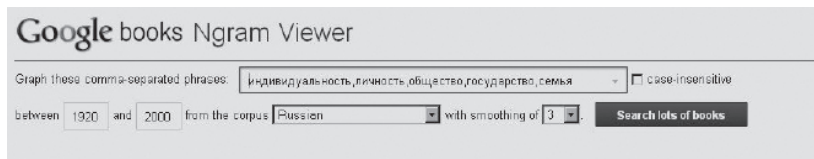


Рис. 9. Область запроса на построение графика.

Graph these case-sensitive comma-separated phrases — построить графики для этих словоформ с учетом регистра, разделять N-граммы запятыми (окно запроса на построение графика);

case-insensitive — при установке флажка в окне система не различает заглавные и строчные буквы;

between and — между ... и... (окно указания временного периода, вводится год начала и конца исследуемого периода);

from the corpus — из корпуса (выбрать из выпадающего меню);

with smoothing — со сглаживанием (выбрать из выпадающего меню);

search lots of books — искать в массивах книг (кнопка команды на поиск и построение графика).

Имеется возможность при формулировке условий поиска задавать распознавание *заглавных и строчных букв* (case sensitive), или игнорировать различие между ними.

В системе нет *грамматической нормализации* лексических единиц, иначе говоря, поиск лексической единицы (слова или словосочетания) и построение графиков частоты ее встречаемости осуществляется для заданной словоформы.

Система предусматривает использование пользовательских тегов для модификации условий построения графиков.

Теги частей речи

Теги этой группы могут применяться изолированно, в этом случае показывается частота употребления данной части речи, а также

могут присоединяться к какому-либо знаменательному слову (больной_NOUN, больной_Adj — см. рис. 10).

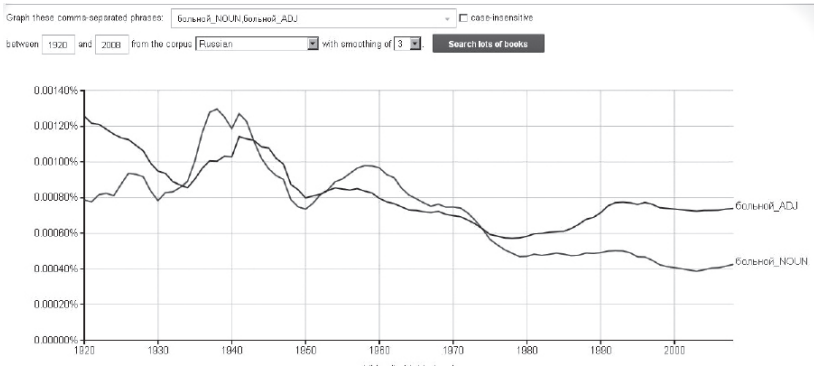


Рис. 10. График встречаемости слова «больной» как существительного и как прилагательного.

Среди прочих тегов имеется тэг_INF (*Inflections*), по которому строятся кривые для всех форм словоизменительной парадигмы данного слова (рис. 11). Следует, однако, отметить, что данная функция работает не всегда корректно, по крайней мере, для русского языка.

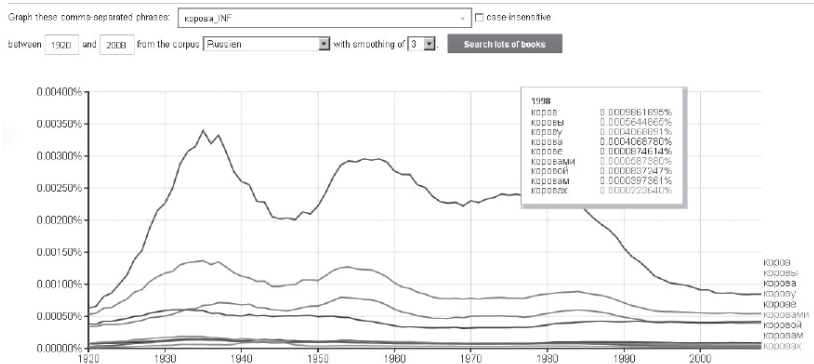


Рис. 11. Частота встречаемости форм словоизменительной парадигмы существительного «корова».

Существуют также теги позиционирования слов (начало и конец предложения, а также поиск глагола, выполняющего роль основного предиката в предложении).

Представляют интерес теги выбора корпусов. В случае их задания система строит графики по разным корпусам одновременно (рис 12).

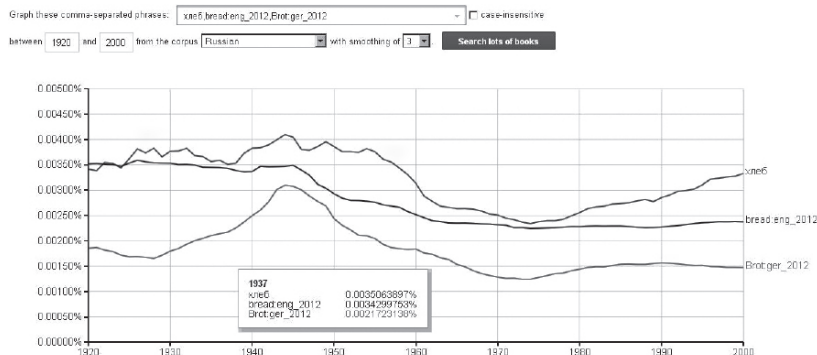


Рис. 12. График встречаемости слова «хлеб» и его английского и немецкого эквивалентов в русском, английском и немецком корпусах.

С октября 2013 года введен тег контекста — «подстановочный знак» * (wildcard). Ввод его через пробел после N-граммы (рис.13) или до неё позволяет строить график встречаемости десяти наиболее частотных сочетаний N-граммы и слова следующего за нею или ей предшествующего.

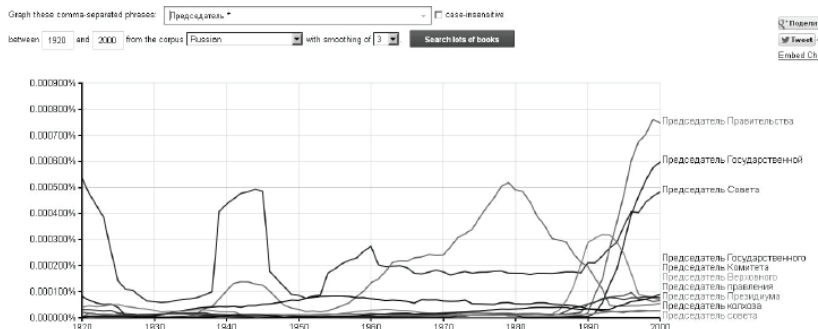


Рис. 13. Кривые встречаемости десяти биграмм с первым словом «Председатель» (использование подстановочного знака после N-граммы).

При задании запроса можно задавать *операции над кривыми графиков*.

Суммирование (сложение) кривых. Операция позволяет суммировать значения каждой точки по оси ординат двух или более кривых. Для осуществления операции поисковые слова вводятся в окно через знак +, например: *лошадь + лошади + лошадей*.

Вычитание кривых. Операция позволяет вычитать из значения каждой точки кривой по оси ординат, значение точки другой кривой для той же позиции по оси абсцисс. С помощью этой операции можно представить, насколько частота встречаемости одной N-граммы больше (меньше) другой, и как это различие менялось во времени. Для осуществления операции поисковые слова вводятся в окно через знак - (дефис), а выражение необходимо брать в круглые скобки, например: *(вежливость-корректность)*. При этой операции вся кривая или ее часть может находиться в области отрицательных значений (см. рис. 14).



Рис. 14. Вычитание значений кривой «корректность» из значений кривой «вежливость».

Есть также умножение и деление графиков. Операция умножения позволяет умножать на число n значения всех точек графика (например, *лемматизация*100*). Данная операция позволяет сделать сопоставимым вид кривых, значения которых отличаются на несколько порядков.

Старая орфография русского языка в Google Books NGram Viewer

Проблема, имеющая особую важность в диахронических исследованиях русского языка – это представление русских текстов в графике и орфографии, действовавшей до 1918 г. включительно. Наличие таких текстов очень ценно для исследования русской лексики. Однако существующие корпуса русского языка, как правило, все тексты дают в современной орфографии. Отсутствие корпусов со старой оригинальной орфографией не позволяет изучить эту сторону языка. К чести НКРЯ, следует сказать, что сейчас там появился исторический подкорпус объемом более 7 млн словоупотреблений.

Тексты Google Books получены посредством оцифровки и распознавания оригинальных печатных изданий. Поэтому в базе данных Google Books тексты книг, изданных до 1919 года (в определенных случаях более поздних изданиях), представлены в старой системе письма, что дает возможность разнообразных исследований. Так, слова с написанием в старой орфографии твердого знака в конце слова можно обрабатывать в двух формах, с твердым знаком и без него (*Бог+Богъ+бог+богъ*), получая на графике суммарную кривую.

Однако другие знаки старой системы русского письма, такие как *і* (код Unicode – 0456), *ѣ* (код Unicode – 0463), *ѳ* (код Unicode – 0473) и соответствующие заглавные буквы, обрабатываемые поисковой системой Google Books, в Google Books NGram Viewer не распознаются. Так, поиск триграммы *Федоръ Михайловичъ Достоевскій* результатов не дает. Но поскольку собственно поисковая система Google осуществляет полноценный поиск по старой русской орфографии, хочется верить, что и в Ngram Viewer скоро эта проблема будет решена.

6.2. Сервис НКРЯ «Графики»

С 2012 г. сервис, аналогичный Google Books NGram Viewer, появился и в России. Текстовый материал НКРЯ по хронологическому материалу также может быть использован для диахронических исследований. Основной массив текстов, собранных в НКРЯ, охватывает период в 200 лет, поэтому он наиболее приспособлен для изучения коротких (несколько десятилетий) и средних (1–2 столетия) языковых изменений. Объем корпуса позволяет изучать вариативность и изменчивость достаточно частотных языковых явлений, а также получать надежные результаты по следующим направлениям:

1) изучение морфологических вариантов имен, глаголов и т. д. и их эволюции;

2) исследование словообразовательных вариантов и связанной с ними проблемы паронимов, продуктивности словообразовательных моделей и словообразовательных средств;

3) исследование изменения вариантов управления, согласования и примыкания;

4) исследование лексической вариативности.

Однако в ряде случаев, например, при изучении фразеологизмов и других устойчивых сочетаний, сервис НКРЯ дает картину мало репрезентативную по причине недостаточного объема данных.

Сервис работает на текстах основного корпуса НКРЯ и называется «Графики». Функционально сервис «Графики» подобен сервису Google Books Ngram Viewer. Он показывает хронологическое распределение заданных и найденных лексических единиц (словоформ, словосочетаний) в основном корпусе НКРЯ. Вход в этот сервис возможен как со страницы с результатами поиска по произвольному запросу к основному корпусу (ссылка *Распределение по годам*), так и из главного меню (рис. 15).

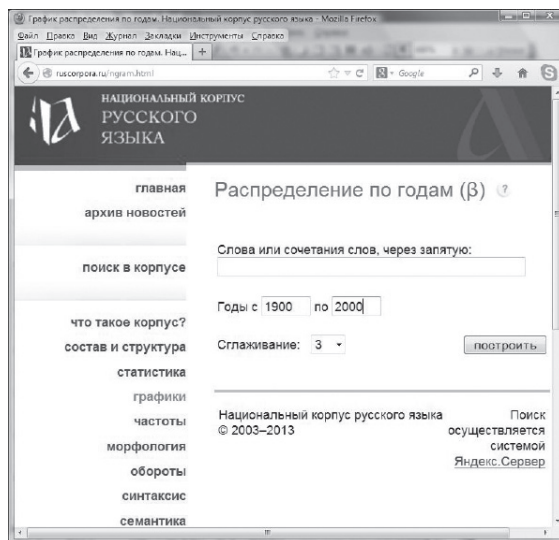


Рис. 15. Интерфейс сервиса «Графики»

Можно задать временные границы, например, с 1930 по 1960 г. Нажав на кнопку «Построить», мы получим график (рис. 16), где каждому элементу запроса (в нашем примере Черчилль, Рузвельт, Франко) соответствует линия своего цвета (см. легенду в правом верхнем углу).

По вертикальной оси графика откладывается относительная частота употреблений данной лексической единицы (в *ipm*). Подводя мышку к любой точке на линии, мы увидим относительную частоту употребления за определенный год (*ipm*). Сглаживание графиков позволяет увидеть общую тенденцию за случайными колебаниями частот. Например, сглаживание в 10 лет усредняет частоту слова с учетом предшествующих и последующих 5 лет, т.е. для данного года берется средняя частота употребления за 11 лет.

Имеется возможность показать таблицы с абсолютными и относительными частотами употреблений за каждый год. Из таблиц гиперссылки позволяют перейти к просмотру примеров из корпуса по годам.

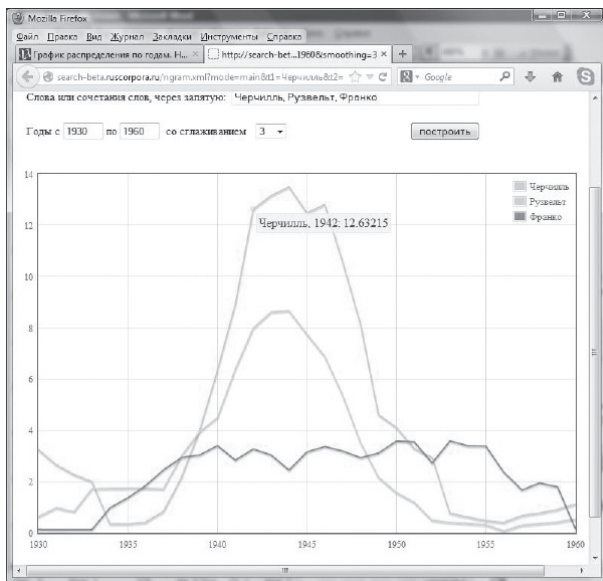


Рис. 16. График встречаемости имен Черчилль, Рузвельт, Франко в текстах, опубликованных с 1930 по 1960 гг.

Из сервиса «Графики» можно перейти на сервис Google Ngram Viewer, работающий на русскоязычной коллекции текстов Google Books. Однако при сходной идеологии, формулы подсчета относительной частоты в сервисах Национального корпуса и Google Ngram Viewer несколько отличаются.

Было бы интересно провести сравнительные исследования результатов по одним и тем же запросам, получаемым по текстам корпуса Google Books Ngram Viewer и НКРЯ.

Литература

Алексеев В. А., Алексеева Е. Л., Касьяненко С. Е. Грамматическая разметка в корпусе СКАТ // Труды международной конференции «Корпусная лингвистика — 2011». СПб, 2011. С. 69–73.

Алексеева Е. Л., Лаврентьев А. М., Азарова И. В., Захарова Л. А. Разметка корпуса древнерусских текстов // Труды международной конференции «Корпусная лингвистика 2004». 11–14 октября 2004 г. СПб., 2004. С. 16–24.

Алексеева Е. Л. Синтаксическая разметка корпуса древнерусских агиографических текстов СКАТ // Структурная и прикладная лингвистика. Выпуск 10. СПб, 2014. С. 345–351.

Андрющенко В. М. Концепция и архитектура Машинного фонда русского языка. Москва: Наука, 1989.

Асиновский А. С., Богданова Н. В., Степанова С. Б., Шерстинова Т. Ю., Маркасова Е. В., Супрунова А. В. Звуковой корпус русского языка «Один речевой день»: пути пополнения и первые результаты исследования // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог», № 9 (16). М.: РГГУ, 2010. С. 41–47.

Богданова С. Ю. О разработке учебного мультимодального корпуса текстов // Материалы XLII международной филологической конференции. Секция прикладной и математической лингвистики. (С. Петербург, 14–19 марта 2013 г.). СПб.: СПбГУ, 2013. С. 27–32.

Бондарко Л. В. Фонетический фонд современного русского языка // Бюллетень фонетического фонда русского языка. 1988. № 1.

Гарабик Р., Захаров В. П. Параллельный Русско-Словацкий Кор-

пус // Труды международной конференции «Корпусная лингвистика–2006». СПб.: Изд-во С. Петерб. ун-та, 2006. С. 81–88.

Гвишиани Н. Б. Практикум по корпусной лингвистике: Учеб. пособие по английскому языку. М.: Высшая школа, 2008.

Герд А. С., Алексеева Е. Л., Азарова И. В., Захарова Л. А. Электронный корпус текстов по памятникам древнерусской агиографической литературы // НТИ. Сер. 2. Вып. 9. 2004. С. 16–20.

Гришина Е. А. Два новых проекта для Национального корпуса: мультимедийный подкорпус и подкорпус названий // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005. С. 233–250.

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

Гришина Е. А. Устная речь в Национальном корпусе русского языка // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005. С. 94–110.

Гришина Е. А., Савчук С. О. Корпус устных текстов в НКРЯ: состав и структура // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 129–149.

Грудева Е. В. Корпусная лингвистика: Учебное пособие. 2-е изд., стереотип. М.: Флинта, 2012.

Добровольский Д. О. Корпус параллельных текстов и литературный перевод // НТИ сер.2, № 10, 2003. С. 13–18.

Добровольский Д. О. Корпус параллельных текстов как инструмент анализа литературного перевода // Компьютерная лингвистика и интеллектуальные технологии. М.: Наука, 2003. С. 126–131.

Добровольский Д. О., Кретов А. А., Шаров С. А. Корпус параллельных текстов: архитектура и возможности использования // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005. С. 263–296.

Ершов А. П. К методологии построения диалоговых систем: феномен деловой прозы // Вопросы кибернетики. Общение с ЭВМ на естественном языке. М., 1982.

Засорина Л. Н. (ред.). Частотный словарь русского языка. М.: 1977.

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

Захаров В. П., Богданова С. Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика» (1-е изд., Иркутск: ИГЛУ, 2011; 2-е изд., перераб. и дополн., СПб.: СПбГУ, Филологический факультет, 2013).

Захаров В. П., Масевич А. Ц. Диахронические исследования на основе корпуса русских текстов Google Books Ngram Viewer. // Структурная и прикладная лингвистика. Выпуск 10. СПб., 2014. С. 303–330.

Камишилова О. Н. Учебный корпус текстов: потенциал, состав, структура. СПб.: ООО «Книжный Дом», 2012.

Кибрик А. А., Подлесская В. И. Рассказы о сновидениях. Корпусное исследование устного дискурса. М., 2009.

Компьютерная лингвистика и интеллектуальные технологии: Материалы ежегодной Международной конференции «Диалог» (2000–2014).

Котов А. А., Гонкало О. С. Русскоязычный эмоциональный корпус // Труды международной конференции «Корпусная лингвистика-2011». Санкт-Петербург. СПб.: СПбГУ. Филологический факультет. 2011. С. 211–216.

Кустова Г. И., Ляшевская О. Н., Падучева Е. В., Рахилина Е. В. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005. С. 155–174.

Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.

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

Митрофанова О. А., Емельянова А. В., Ильина А. И., Кулажко С. А., Михайлова В. Д., Якуба Н. М. Лексико-грамматические особенности текстов специальных корпусов разной тематики // Сборник материалов по итогам XLIII Международной филологической конференции. Секция прикладной и математической лингвистики. СПб., 2014.

Михайлов М. Н. Параллельные корпуса художественных текстов: принципы составления и возможности применения в лингвистических и переводоведческих исследованиях. Тампере: Тамперский университет, 2003.

Мустайоки А., Копотев М. В., Гурин Г. Б., Саломатина М. С. Принципы синтаксической разметки Хельсинского аннотированного корпуса русских текстов ХАНКО // Труды международной конференции «MegaLing`2005. Прикладная лингвистика в поиске новых путей». СПб., 2005. С. 90–95.

Национальный корпус русского языка: 2003–2005. Сборник статей. М.: Индрик, 2005.

Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009.

Николаева Ю. В. Сегментация устного нарратива и изобразительные жесты: кинетические признаки границ и связей между сегментами дискурса // Компьютерная лингвистика и интеллектуальные технологии: Материалы ежегодной Международной конференции «Диалог». М.: РГГУ, 2009. С. 340–345.

Резникова Т. И. Славянская корпусная лингвистика: современное состояние ресурсов // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 402–461.

Резникова Т. И., Копотев М. В. Лингвистически аннотированные корпуса русского языка (обзор общедоступных ресурсов) // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005. С. 31–61.

Труды международной конференции «Корпусная лингвистика» (2004, 2006, 2008, 2011, 2013). Санкт-Петербург. СПб.: СПбГУ. Филологический факультет.

Шерстинова Т. Ю., Степанова С. Б., Рыко А. И. Система аннотирования в звуковом корпусе русского языка «Один речевой день» // Мат-лы XXXVIII международной филологической конференции. Секция: «Формальные методы анализа русской речи». СПб.: СПбГУ, 2009. С. 66–75.

Asinovsky A., Bogdanova N., Rusakova M., Stepanova S., Ryko A., Sherstinova T. The ORD Speech Corpus of Russian Everyday Communication «One Speaker's Day»: Creation Principles and Annotation // Lecture Notes in Computer Science, 2009. Vol. Text, Speech and Dialogue, № 5729/2009. P. 250–257.

Benko V. Aranea: Yet Another Family of (Comparable) Web Corpora, In: Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno,

Czech Republic, September 8–12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, pp. 257–264, ISBN: 978–3-319–10815-5.

Čermák F, Rosen A. The case of InterCorp, a multilingual parallel corpus. International In: Journal of Corpus Linguistics. 2012. Vol. 13, no. 3, pp. 411–427.

Kilgarriff A., Rychly P., Smrz P., Tugwell D. The Sketch Engine // Proceedings of the XIth Euralex International Congress. Lorient: Universite de Bretagne-Sud, 2004, pp. 105–116.

Michel J.B. et al. Quantitative Analysis of Culture Using Millions of Digitized Books science. Science 331, 176 (2011); DOI 1126/Science. 1199644 URL: <http://www.sciencemag.org/content/331/6014/176.full.html>

Steinberger Ralf, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski & Signe Gilbro. An overview of the European Union's highly multilingual parallel corpora. Language Resources and Evaluation Journal (LRE) 2014. DOI: 10.1007/s10579–014–9277–0.

Waldenfels Ruprecht, von. Compiling a parallel corpus of Slavic languages. In: Bernhard Brehmer, Vladislava Ždanova, Rafał Zimny (Hrsg.) Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9. München, 2006. S.123–138.

Waldenfels Ruprecht, von. Recent developments in ParaSol: Breadth for depth and XSLT based web concordancing with CWB. In: Daniela M., and Garabík, R. (eds.), Natural Language Processing, Multilinguality. Proceedings of Slovko 2011 (Modra, Slovakia, 20–21 October 2011). Bratislava, 2011, pp. 156–162.

Zakharov V. Corpora of the Russian Language // Text, Speech and Dialogue: Proceedings of the 16th International Conference, TSD 2013, Plzen, Czech Republic, September 1–5, 2013. Ivan Habernal, Václav Matoušek (Eds.). Springer-Verlag, Berlin Heidelberg, 2013, pp. 113. (Lecture Notes in Artificial Intelligence, 8082).

V. P. Zakharov

¹*Saint-Petersburg State University*

²*Institute for Linguistic Studies*

(Russia, Saint-Petersburg)

vz1311@yandex.ru

CORPORA OF THE RUSSIAN LANGUAGE

The paper describes corpora of the Russian language and the state of the art of Russian corpus linguistics. The main attention is paid to the Russian National Corpus (RNC) as the most popular one among linguists for both being the most well known and the opportunities which it presents. The author regards the subcorpora within the RNC, semantic annotation and Charts service of the RNC in comparison with Google Books Ngram Viewer. The article also presents a large number of other text corpora of Russian, among them Helsinki Annotated Corpus (HANCO), Leeds University corpora, Sketch Engine corpora and so on, speech corpora of Russian, parallel corpora, diachronic corpora and specialized corpora. Corpora of the Russian language provide corpus-based studies of both oral speech and written language using synchronic and diachronic approaches and allow us to study linguistic phenomena in typological, sociological, culturological aspects.

Key words: Russian corpus linguistics, corpora, the Russian language, specialized corpora.

References

Alekseev V. A., Alekseeva E. L., Kas'yanenko S. E. [Grammatical Markup of the "SKAT" Corpus]. *Trudy mezhdunarodnoi konferentsii "Korpusnaya lingvistika — 2011"* [Proceedings of the International Conference "Corpus Linguistics — 2011"]. St. Petersburg, 2011, pp. 69–73. (In Russ.)

Alekseeva E. L., Lavrent'ev A. M., Azarova I. V., Zaharova L. A. [Markup of the Old Russian Language Corpus]. *Trudy mezhdunarodnoi konferentsii "Korpusnaya lingvistika – 2004"* [Proceedings of the

International Conference “Corpus Linguistics – 2004”]. 11-14 October, 2004. St. Petersburg, 2004, pp. 16–24. (In Russ.)

Alekseeva E. L. [Syntactic Markup of the Corpus of Old Russian Hagiographical Texts “SKAT”]. *Strukturnaya i prikladnaya lingvistika* [Structural and applied linguistics]. Issue 10. St. Petersburg, 2014, pp. 345–351. (In Russ.)

Andryushchenko V. M. *Kontsepsiya i arkhitektura Mashinnogo fonda russkogo yazyka* [Conception and Architecture of the Foundation for the Russian Language Machine Analysis]. Moscow, Nauka Publ., 1989. (In Russ.)

Asinovsky A. S., Bogdanova N. V., Stepanova S. B., Sherstinova T. Ju., Markasova E. V., Suprunova A. V. [“One Speaker's Day” Sound Corpus of the Russian Language: Ways of Stocking and First Research Results]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoi konferentsii “Dialog – 2010”* [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference “Dialog–2010”]. Moscow, RSUH Publ., 2010, pp. 41–47. (In Russ.)

Asinovsky A., Bogdanova N., Rusakova M., Stepanova S., Ryko A., Sherstinova T. The ORD Speech Corpus of Russian Everyday Communication “One Speaker's Day”: Creation Principles and Annotation. *Lecture Notes in Computer Science, 2009*. Vol. Text, Speech and Dialogue, № 5729/2009, pp. 250–257.

Benko V. Aranea: Yet Another Family of (Comparable) Web Corpora. Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): *Text, Speech and Dialogue. 17th International Conference*, TSD 2014, Brno, Czech Republic, September 8–12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, pp. 257–264.

Bogdanova S. Ju. [On development of the educational multi-language corpus]. *Materialy XLII mezhdunarodnoi filologicheskoi konferentsii. Sektsiya prikladnoi i matematicheskoi lingvistiki* [Proceedings of the XLII International Philological Conference. Group for Applied and Mathematical Linguistic]. (S. Peterburg, 14-19 March, 2013). St. Petersburg, St. Petersburg St. Univ. Publ., 2013, pp. 27–32. (In Russ.)

Bondarko L. V. [Phonetic Foundation of the Modern Russian Language]. *Byulleten' foneticheskogo fonda russkogo yazyka*, 1988, no. 1. (In Russ.)

Čermák F., Rosen A. The case of InterCorp, a multilingual parallel

corpus. International In: *Journal of Corpus Linguistics*, 2012, Vol. 13, no. 3, pp. 411–427.

Dobrovol'skij D. O. [Corpus of Parallel Texts and Literary Translation]. *Nauchno-tehnicheskaya informatsiya*, Ser. 2, 2003, no. 10, pp. 13–18. (In Russ.)

Dobrovol'skij D. O. [Corpus of Parallel Texts as an Instrument for Literary Translation Analysis]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii* [Computational Linguistics and Intellectual Technologies]. Moscow, Nauka Publ., 2003, pp. 126–131. (In Russ.)

Dobrovol'skij D. O., Kretov A. A., Sharov S. A. [Corpus of Parallel Texts: Architecture and Usage Possibilities]. *Natsional'nyi korpus russkogo yazyka: 2003–2005* [The National Corpus of the Russian Language: 2003–2005]. Moscow, Indrik Publ., 2005, pp. 263–296. (In Russ.)

Ershov A. P. [On the Methodology of Dialog Systems Building: Phenomenon of Business Texts]. *Voprosy kibernetiki. Obshchenie s EVM na estestvennom yazyke* [Cybernetic Questions. Communication with Computer Using Natural Language]. Moscow, 1982. (In Russ.)

Garabik R., Zaharov V. P. [Parallel Russian — Slovak Corpus]. *Trudy mezhdunarodnoi konferentsii “Korpusnaya lingvistika – 2006”* [Proceedings of the International Conference “Corpus Linguistics — 2006”]. St. Petersburg, St. Petersburg St. Univ. Publ., 2006, pp. 81–88. (In Russ.)

Gerd A. S., Alekseeva E. L., Azarova I. V., Zaharova L. A. [Electronic Corpus of the Old Russian Hagiographic Texts]. *Nauchno-tehnicheskaya informatsiya*, Ser. 2. 2004, Issue 9, pp. 16–20. (In Russ.)

Grishina E. A. [Multi-Media Corpus of the Russian Language (MURKO): Problems with Annotation]. *Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy* [The National Corpus of the Russian Language: 2006–2008. New Results and Prospective]. St. Petersburg, Nestor-Istoriya Publ., 2009, pp. 175–214. (In Russ.)

Grishina E. A. [Two New Projects for the National Corpus of the Russian Language: Multi-Media Subcorpus and Subcorpus of Titles]. *Natsional'nyi korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy* [The National Corpus of the Russian Language: 2003–2005. Results and Prospective]. Moscow, Indrik Publ., 2005, pp. 233–250. (In Russ.)

Grishina E. A., Savchuk S. O. [Corpus of the Spoken Language in the National Corpus of the Russian Language: content and structure]. *Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i*

perspektivy [National Corpus of the Russian Language: 2006–2008. New Results and Prospective]. St. Petersburg, Nestor-Istoriya Publ., 2009, pp. 129–149. (In Russ.)

Grishina E. A. [Spoken Language in the National Corpus of the Russian Language]. *Natsional'nyi korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy* [National Corpus of the Russian Language: 2003–2005. Results and Prospective]. Moscow, Indrik Publ., 2005, pp. 94–110. (In Russ.)

Grudeva E. V. *Korpusnaya lingvistika: Uchebnoe posobie* [Corpus Linguistics: Educational Manual]. 2nd ed. Moscow, Flinta Publ., 2012.

Gvishiani N. B. *Praktikum po korpusnoi lingvistike: Ucheb. posobie po angliiskomu yazyku* [Workshop on Corpus Linguistics: Educational Manual on the English Language]. Moscow, Vysshaya shkola Publ., 2008.

Kamshilova O. N. *Uchebnyi korpus tekstov: potentsial, sostav, struktura* [Educational Corpus: Potential, Content, and Structure]. St. Petersburg, “Knizhnyi Dom” Publ., 2012.

Kibrik A. A., Podlesskaya V. I. *Rasskazy o snovideniyakh. Korpusnoe issledovanie ustnogo diskursa* [Description of Dreams: Corpus Research of Spoken Discourse]. Moscow, 2009.

Kilgarriff A., Rychly P., Smrz P., Tugwell D. The Sketch Engine. *Proceedings of the XIth Euralex International Congress*. Lorient, Universite de Bretagne-Sud, 2004, pp. 105–116.

Komp'yuternaya lingvistika i intellektual'nye tekhnologii: *Materialy ezhegodnoi Mezhdunarodnoi konferentsii “Dialog”* [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference “Dialog”] (2000–2014).

Kotov A. A., Gopkalo O. S. *Russkoyazychnyi emotsional'nyi korpus* [Emotional Corpus for the Russian Language]. *Trudy mezhdunarodnoi konferentsii “Korpusnaya lingvistika – 2011”* [Proceedings of the International Conference “Corpus Linguistics – 2011”]. St. Petersburg, St. Petersburg St. Univ. Publ., 2011, pp. 211–216. (In Russ.)

Kustova G. I., Lyashevskaya O. N., Paducheva E. V., Rahilina E. V. [Semantic Markup in the National Corpus of the Russian Language: Principles, Problems, Prospective]. *Natsional'nyi korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy* [National Corpus of the Russian Language: 2003–2005. Results and Prospective]. Moscow, Indrik Publ., 2005, pp. 155–174. (In Russ.)

Lyashevskaja O. N., Sharov S. A. *Chastotnyi slovar' sovremennogo russkogo yazyka (na materialakh Natsional'nogo korpusa russkogo yazyka)* [Frequency Dictionary of the Modern Russian Language (Based on the Data of the National Corpus of the Russian Language)]. Moscow, Azbukovnik Publ., 2009.

Mashinnyi fond russkogo yazyka: idei i suzhdeniya [Computerized Foundation for the Russian Language: Ideas and Opinions]. Moscow, Nauka Publ., 1986.

Michel J. B. et al. Quantitative Analysis of Culture Using Millions of Digitized Books science. Science 331, 176 (2011); DOI 1126/Science. 1199644. Available at: URL: <http://www.sciencemag.org/content/331/6014/176.full.html> (accessed 13.06.2015)

Mihajlov M. N. *Parallel'nye korpusa khudozhestvennykh tekstov: printsipy sostavleniya i vozmozhnosti primeneniya v lingvisticheskikh i perevodovedcheskikh issledovaniyakh* [Parallel Corpora of Fiction: Principles of Building and Possibilities of Usage through Researches on Linguistics and Translation]. Tampere, 2003. (In Russ.)

Mitrofanova O. A., Emel'yanova A. V., Il'ina A. I., Kulazhko S. A., Mikhailova V. D., Yakuba N. M. [Lexical and Grammatical Features of the Corpus of the Technical Texts of Different Aspects]. *Materialy XLII mezhdunarodnoi filologicheskoi konferentsii. Sektsiya prikladnoi i matematicheskoi lingvistiki* [Proceedings of the XLII International Philological Conference. Group for Applied and Mathematical Linguistics]. (St. Petersburg, 11–16 March, 2014). St. Petersburg, St. Petersburg St. Univ. Publ., 2014. (In Russ.)

Mustajoki A., Kopotev M. V., Gurin G. B., Salomatina M. S. [Principles of Syntactic Markup of the HANCO (Helsinki Annotated Corpus of Russian Texts)]. *Trudy mezhdunarodnoi konferentsii "MegaLing'2005. Prikladnaya lingvistika v poiske novykh putei"* [Proceedings of the International Conference "MegaLing'2005. Applied Linguistics: Looking for New Ways"]. St. Petersburg, 2005, pp. 90–95. (In Russ.)

Natsional'nyi korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy [Russian National Corpus: 2003–2005. Results and Prospective]. Moscow, Indrik Publ., 2005.

Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy. [Russian National Corpus: 2006–2008. New Results and Prospective]. St. Petersburg, Nestor-Istoriya Publ., 2009.

Nikolaeva Yu. V. [Segmentation of Spoken Narrative and Pictorial

Gestures: Kinetic Indication of Borders and Bounds between Segments of Discourse]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Materialy ezhegodnoi Mezhdunarodnoi konferentsii "Dialog–2009"* [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog–2009"]. Moscow, RSUH Publ., 2009, pp. 340–345. (In Russ.)

Reznikova T. I. [Slavic Corpus Linguistics: Modern State of Recourses]. *Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy* [Russian National Corpus: 2006–2008. New Results and Prospective]. St. Petersburg, Nestor-Istoriya Publ., 2009, pp. 402–461. (In Russ.)

Reznikova T. I., Kopotev M. V. [Linguistically Annotated Corpora of the Russian Language (Review of the Public Resources)]. *Natsional'nyi korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy* [Russian National Corpus: 2003–2005. Results and Prospective]. Moscow, Indrik Publ., 2005, pp. 31–61. (In Russ.)

Sherstinova T. Yu., Stepanova S. B., Ryko A. I. [Markup System of the ["One day of speech" Sound Corpus of the Russian Language]. *Materialy XXXVIII mezhdunarodnoi filologicheskoi konferentsii. Sektsiya "Formal'nye metody analiza russkoi rechi"* [Proceeding of the XXXVIII International Philological Conference. "Formal Methods of Russian Speech Analysis" Group]. St. Petersburg, St. Petersburg St. Univ. Publ., 2009, pp. 66–75. (In Russ.)

Steinberger Ralf, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski & Signe Gilbro. An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation Journal (LRE)* 2014. DOI: 10.1007/s10579-014-9277-0.

Trudy mezhdunarodnoi konferentsii "Korpusnaya lingvistika" (2004, 2006, 2008, 2011, 2013) [Proceedings of the International Conference "Corpus Linguistics" (2004, 2006, 2008, 2011, 2013)]. St. Petersburg, St. Petersburg St. Univ., Philological faculty.

Waldenfels Ruprecht, von. Compiling a parallel corpus of Slavic languages. Bernhard Brehmer, Vladislava Ždanova, Rafał Zimny (Hrsg.) *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV)* 9. München, 2006, pp.123–138.

Waldenfels Ruprecht, von. Recent developments in ParaSol: Breadth for depth and XSLT based web concordancing with CWB. Daniela M.,

and Garabík, R. (eds.), *Natural Language Processing, Multilinguality. Proceedings of Slovo 2011* (Modra, Slovakia, 20–21 October 2011). Bratislava, 2011, pp. 156–162.

Zakharov V. Corpora of the Russian Language. *Text, Speech and Dialogue: Proceedings of the 16th International Conference, TSD 2013*, Plzen, Czech Republic, September 1–5, 2013. Ivan Habernal, Václav Matoušek (Eds.). Springer-Verlag, Berlin Heidelberg, 2013, pp. 1–13. (Lecture Notes in Artificial Intelligence, 8082).

Zakharov V. P. *Korpusnaya lingvistika: Uchebno-metod. posobie* [Corpus Linguistics: Educational and Methodological Manual]. St. Petersburg, 2005.

Zakharov V. P., Bogdanova S. Ju. *Korpusnaya lingvistika: Uchebnik dlja studentov napravlenija "Lingvistika"* [Corpus Linguistics: Manual for Students of Linguistic Departments]. (1st ed., Irkutsk: IGLU, 2011; 2nd ed., St. Petersburg, St. Petersburg St. Univ., Philological faculty, 2013).

Zakharov V. P., Masevich A. Ts. [Diachronical Research Based on Google Books Ngram Viewer Corpus of Russian texts]. *Strukturnaya i prikladnaya lingvistika* [Structural and Applied Linguistics]. Issue 10, St. Petersburg, 2014, pp. 303–330. (In Russ.)

Zasorina L. N. (Ed.). *Chastotnyi slovar' russkogo yazyka* [Frequency Dictionary of the Russian Language]. Moscow, 1977.