МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ
ФГБОУ ВПО «Санкт-Петербургский государственный университет»

В.П. Захаров, С.Ю. Богданова

КОРПУСНАЯ ЛИНГВИСТИКА

Рекомендован Учебно-методическим объединением по образованию в области лингвистики Министерства образования и науки Российской Федерации в качестве учебника для студентов, обучающихся по направлению подготовки бакалавров и магистров 035700 «Лингвистика»

Санкт-Петербург
2013

УДК 81.32
ББК 81.1-923
З-38

Рецензенты:
Докт. филол. наук, *С.А. Крылов* (ИВ РАН)
Докт. техн. наук, проф. *В.Ш. Рубашкин* (СПбГУ)

Печатается по постановлению
Редакционно-издательского совета
Филологического факультета СПбГУ

Захаров В.П., Богданова С.Ю.

З-38 Корпусная лингвистика: Учебник для студентов направления «Лингвистика». 2-е изд., перераб. и дополн., – СПб.: СПбГУ. РИО. Филологический факультет, 2013. – 148 с.

Учебник знакомит с концепциями корпусной лингвистики, дает возможность освоить основы корпусных технологий, приобрести навыки работы с корпусами, определить место дисциплины и собственно корпусов в ряду информационных технологий. Учебник базируется на исследовательской и преподавательской деятельности авторов.

Предназначен для студентов, магистрантов и аспирантов филологических специальностей.

УДК 81.32
ББК 81.1-923

ПРЕДИСЛОВИЕ

Предлагаемый вашему вниманию учебник является своего рода обобщением многочисленных разрозненных материалов, опубликованных за последние два десятилетия в России и за рубежом, которые легли в основу лекционных курсов по дисциплине «Корпусная лингвистика», читаемых кандидатом филологических наук, доцентом Виктором Павловичем Захаровым в Санкт-Петербургском государственном университете и доктором филологических наук, профессором Светланой Юрьевной Богдановой в Иркутском государственном лингвистическом университете. Материал, представленный в учебном издании, может также быть использован в курсах лекций по дисциплинам «Информационные и коммуникационные технологии в науке и образовании», «Основы прикладной лингвистики», «Компьютерные методы в лингвистических исследованиях» и др.

Цель учебника – познакомить студентов с концепциями корпусной лингвистики, помочь им освоить основы корпусных технологий, приобрести навыки работы с корпусами, определить место дисциплины и собственно корпусов в ряду информационно-лингвистических технологий.

Задачи учебника:

- ознакомить студентов с новой парадигмой в лингвистических исследованиях;
- ознакомить студентов с историей корпусных исследований;
- ознакомить студентов с языковыми и программными средствами корпусной лингвистики;
- сформировать у студентов навыки работы с программными средствами и информационными ресурсами корпусной лингвистики;
- ознакомить студентов с конкретными лингвистическими исследованиями, основанными на корпусных данных.

Учебник состоит из трех частей. Первая часть «Введение в корпусную лингвистику» знакомит с основными понятиями и терминами корпусной лингвистики, историей ее становления как раздела языкознания, ее целями и задачами, типами существующих корпусов. Вторая

часть «Создание корпусов» описывает в общих чертах технологические процессы, связанные с их проектированием, отбором и обработкой языкового материала, способами разметки. Третья часть «Использование корпусов» включает три раздела. Раздел 3.1. «Корпусные менеджеры» посвящен описанию корпусных менеджеров, обеспечивающих поиск в корпусе. Раздел 3.2. «Обзор существующих корпусов различных типов» представляет собой обзор как зарубежных национальных корпусов, так и корпусов русского языка. Раздел 3.3. «Корпусные исследования» посвящен описанию конкретных исследований на базе корпусов разных типов, в нем приводятся результаты научных изысканий и дается их теоретическая интерпретация.

В первую очередь авторы хотят показать, как можно, базируясь на корпусах, работать с реальным языковым материалом быстрее и эффективнее. В этом разделе приведены примеры исследований лишь в нескольких областях лингвистики – лексикографии, грамматике и анализе дискурса. Безусловно, сфера применения корпусных данных в лингвистике значительно шире. В Приложении приведен краткий глоссарий терминов корпусной лингвистики.

Надеемся, что студенты направления «Лингвистика» заинтересуются использованием корпусов, независимо от сферы их научных интересов, а каждый преподаватель найдет в учебнике то, о чем нужно говорить его аудитории.

Авторы выражают искреннюю благодарность заведующему кафедрой математической лингвистики СПбГУ Александру Сергеевичу Герду за критические замечания и рекомендации, сделанные в процессе подготовки учебника.

Часть 1

ВВЕДЕНИЕ В КОРПУСНУЮ ЛИНГВИСТИКУ

1.1. Основные понятия корпусной лингвистики

Корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с применением компьютерных технологий. Под *лингвистическим,* или *языковым, корпусом текстов* понимается большой, представленный в машиночитаемом формате, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач. В настоящее время существует множество определений понятия «корпус». Например, определение, приведенное в учебнике Э. Финегана, гласит: *корпус* – репрезентативное собрание текстов, обычно в машиночитаемом формате, включающее информацию о ситуации, в которой текст был произведен, такую как информация о говорящем, авторе, адресате или аудитории [42].

Википедия определяет корпусы как большие и структурированные наборы текстов (теперь обычно в электронном виде), которые используются для статистического анализа и проверки гипотез, проверки случаев встречаемости или обоснования языковых правил по определенным областям. Т. МакЭнери (T. McEnery) и Э. Вилсон (A. Wilson) дают следующее определение: *корпус* – это собрание языковых фрагментов, отобранных в соответствии с четкими языковыми критериями для использования в качестве модели языка [51].

В приведенных определениях подчеркиваются основные черты современного корпуса текстов – цель («логическая идея»), машиночитаемый формат, репрезентативность как результат особой процедуры отбора, наличие металингвистической информации. Стандартизованное представление словесного материала на машинном носителе позволяет применять стандартные программы его обработки.

Целесообразность создания и смысл использования корпусов определяется следующими предпосылками:

1) достаточно большой (репрезентативный) и сбалансированный объем корпуса гарантирует типичность данных и обеспечивает полноту представления всего спектра языковых явлений;

2) данные разного типа находятся в корпусе в своей естественной контекстной форме, что создает возможность их всестороннего и объективного изучения;

3) однажды созданный и подготовленный массив данных может использоваться многократно, различными исследователями и в различных целях.

В понятие «корпус текстов» входит также система управления текстовыми и лингвистическими данными, которую в последнее время чаще всего называют *корпусным менеджером* (или корпус-менеджером) (*англ.* corpus manager). Это специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления пользователю результатов в удобной форме.

Поиск в корпусе позволяет по любому слову построить *конкорданс* – список всех употреблений данного слова в контексте со ссылками на источник. Корпусы могут использоваться для получения разнообразных справок и статистических данных о языковых и речевых единицах. В частности, на основе корпусов можно получить данные о частоте словоформ, лексем, грамматических категорий, проследить изменение частот и контекстов в различные периоды времени, получить данные о совместной встречаемости лексических единиц и т. д. Представительный массив языковых данных за определенный период позволяет изучать динамику процессов изменения лексического состава языка, проводить анализ лексико-грамматических характеристик в разных жанрах и у разных авторов.

Корпусы призваны служить источником и инструментом многоаспектных лексикографических работ по подготовке разнообразных исторических и современных словарей. Данные корпусов могут быть использованы для построения и уточнения грамматик, а также в процессе обучения языку. Более подробно возможности и примеры использования корпусов в лингвистических исследованиях будут рассмотрены в разд. 3.3.

Сегодня корпусная лингвистика часто понимается как относительно новый подход в лингвистике, который имеет дело с изучением использования языка в «реальной жизни» с помощью компьютеров и

электронных корпусов. Корпусная лингвистика имеет, по крайней мере, две черты, дающие ей основание претендовать на положение самостоятельной дисциплины: 1) характер используемого словесного материала; 2) специфика инструментария.

Если такие разделы лингвистики, как синтаксис, семантика и социолингвистика, имеют целью описание или оценку языковой структуры или языкового использования, то корпусная лингвистика является более широким понятием, методологией, которую можно применить ко многим аспектам языковых исследований. Корпусную лингвистику иногда называют «пучком методов из разных областей лингвистических исследований» [49]. Как метод лингвистического анализа корпусная лингвистика связана также с контрастивными исследованиями, направленными на установление фактов общего и отдельного между языками, диалектами или вариантами языка в ходе их сопоставительного изучения [8]. Многие виды лингвистического анализа наилучшим образом развиваются на прочной и обширной базе эмпирических данных.

Э. Финеган определяет корпусную лингвистику как деятельность, требующуюся для составления и использования корпуса и направленную на исследование естественного употребления языка [42]. В этом определении подчеркивается созидательная направленность корпусной лингвистики. Двойственный характер корпусной лингвистики (нацеленность как на создание, так и на использование корпусов текстов) обусловливается двойственным характером ее *объекта* – корпуса текстов, который, с одной стороны, представляет собой исходный речевой материал для корпусной лингвистики и для других лингвистических дисциплин; с другой стороны, является продуктом корпусной лингвистики.

Можно сказать, что корпусная лингвистика имеет своим *предметом* теоретические основы и практические механизмы создания и использования представительных массивов языковых данных, предназначенных для лингвистических исследований в интересах широкого круга пользователей.

Существует проблема, связанная с терминологией корпусной лингвистики в русском языке (см. Приложение 2), которая пока не установилась в силу следующих причин: ее относительно недавнее происхождение и ее зарождение в США и Великобритании, обусловившее тот факт, что терминология складывалась и продолжает складываться в недрах английского языка. Русские термины, в основном, представляют собой заимствования английских терминов, некоторые из них в других значениях давно существуют в русском языке. Так, русское

слово «корпус» стало многозначным задолго до своего появления в качестве термина корпусной лингвистики. Употребление форм этого существительного в лингвистике является проблематичным, поскольку возможны варианты множественного числа «корпусы» и «корпуса». Для значения «массив», которое имеет место в случае языковых корпусов, именительный падеж множественного числа должен быть «ко́рпусы» и, соответственно, прилагательное должно произноситься с ударением на первом слоге – «ко́рпусный» (Большой толковый словарь русского языка, СПб., 1998). В то же время анализ узуса специалистов пока свидетельствует в пользу форм «корпуса́», «корпусно́й», «корпусна́я», которые используются часто, так что можно, видимо, с осторожностью сказать, что в настоящее время этот вопрос остается открытым. Правила, регламентирующего употребление той или иной формы применительно к корпусной лингвистике, пока нет, хотя, как представляется, победить должен вариант «корпусы», поскольку он отличает терминологическое значение слова от его общеупотребительного значения. В учебнике авторы будут использовать именно этот вариант.

1.2. Направления в лингвистике, предвосхитившие появление корпусной лингвистики: от картотеки к корпусу

Корпусная лингвистика может быть представлена в виде совокупности методов, процедур и ресурсов, имеющих дело с эмпирическими данными в лингвистике. Подъем современной корпусной лингвистики как методологии тесно связан с историей лингвистики как эмпирической науки.

Технологии, которые применяются в корпусной лингвистике, намного старше электронных компьютеров: многие из них коренятся в традиции конца XVIII и XIX вв., когда лингвистика впервые была провозглашена «реальной», или эмпирической, наукой. Из многочисленных областей лингвистических исследований, которые легли в основу корпусной лингвистики, здесь будут рассмотрены три. Использованные в этих трех областях технологии повлияли на развитие современной корпусной лингвистики, и, наоборот, сейчас она существенно меняет «пейзаж» современной лингвистики [49].

1. Историческая лингвистика: изменения в языке и реконструкция (сравнительно-исторический метод). Одно из главных направлений, повлиявших на современную корпусную лингвистику, пришло из сравнительно-исторического языкознания. Это неудивительно, поскольку лингвисты, занимающиеся историческими исследо-

ваниями, всегда использовали тексты или собрания текстов как основные свидетельства. Многие технологии, развитые в XIX в. для реконструкции более древних языков (праязыков) или установления связей между языками, используются и по настоящее время. В индоевропейской традиции изучение языковых изменений и попытки реконструкции зависели от ранних текстов или корпусов (исторических памятников). Я. Гримм и позднее младограмматики поддерживали свои утверждения об истории и грамматике языков цитатами из текстов. Младограмматики в своем манифесте провозгласили, что они провели исследование современного языка, зафиксированного в диалектах (а не только исследование древних текстов), и это также имело огромное значение.

Многие идеи, развиваемые с XIX в., были применены и затем развиты корпусной лингвистикой. Среди первых корпусов, доступных в электронном виде, были и исторические корпусы.

Появление огромного количества текстов, доступных в электронном формате, предоставило возможность лингвистам широко применять в лингвистическом анализе статистические методы, а также разрабатывать и развивать новые методы и модели исследований. Сегодня математически сложные модели языковых изменений могут быть построены на основе электронных корпусов.

2. Написание грамматик, составление словарей и обучение языку. Грамматисты XIX в. иллюстрировали свои утверждения примерами, взятыми из произведений признанных авторов. Например, Г. Пауль в своей немецкой грамматике использовал произведения немецких «классиков» для иллюстрации каждого своего положения – в области фонологии, морфологии и синтаксиса. Сегодня составители грамматик также используют корпусный подход, но теперь корпусы включают не только классику, но и другие типы текстов и позволяют описать язык более адекватно. В частности, большой интерес проявляется сейчас к грамматике устной речи.

В грамматических описаниях языка можно использовать корпусы для получения информации о частотных характеристиках использования разных вариантов, регистров (жанров)[1] и т. д.

[1] Термины «жанр» и «регистр» часто употребляются в литературе по корпусной лингвистике как синонимы, что, как представляется, зависит от предпочтений авторов. Тем не менее попытки «развести» эти термины неоднократно предпринимались (см. Lee D. Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle // Language

Возьмем некоторые ранние примеры корпусного подхода из лексикографии. В середине XVIII в., когда С. Джонсон составлял толковый словарь английского языка (Dictionary of the English language, 1755), он выбирал из книг иллюстративные предложения, которые называл цитатами, чтобы показать на примерах, как слова использовались английскими авторами. Во время чтения С. Джонсон маркировал предложения, контекст которых делал значение слова особенно понятным. Его ассистенты затем выписывали отмеченные предложения на листы бумаги, и С. Джонсон распределял их для составления и иллюстрации словарных статей в словаре. Проект под руководством сэра Джеймса Муррея (Оксфордский словарь английского языка – OED) потребовал тысячи читателей и полвека для составления.

Многие словари мертвых языков давали цитаты из текстов, содержащие слово в контексте. В современной корпусной лингвистике этот метод параллелен по форме конкордансу KWIC (Key Word in Context). Несмотря на то что компьютеры облегчили поиск и классификацию примеров и выделение многословных единиц, идеи использования текстов из корпуса все еще очень схожи с теми, что применялись ранними лексикографами и филологами, не имевшими доступа к компьютерным технологиям.

Традиционные школьные грамматики и учебники часто проиллюстрированы искусственно составленными или отредактированными примерами языкового употребления. В будущем они мало чем смогут помочь студентам, которые рано или поздно сталкиваются с реальными языковыми данными в своих заданиях или в реальном общении. В этом отношении корпусы как источники эмпирических данных играют важную роль в лингводидактике. При обучении языку корпусы обеспечивают источник для пробуждения у студентов интереса и вовлечение их в самостоятельное изучение аутентичного языкового использования. Важное применение корпусных данных – технологии Computer-Assisted Language Learning (CALL), где основанное на корпусе про-

---

Learning & Technology. Vol. 5, No. 3. September 2001. P. 37–72 (http://llt.msu.edu/vol5num3/lee/default.html), где «жанры» определяются как группы текстов, собранных и скомпилированных для корпусов или корпусных исследований, которые понимаются как *категории* текстов и ассоциируются с типичными конфигурациями власти, идеологии, общественных целей. «Регистры», в свою очередь, акцентируют внимание на параметрах ситуации языкового употребления и имеют естественную ассоциацию с определенными *лингвистическими чертами,* в то время как общественная оценка контекстуального использования уже задана.

граммное обеспечение используется для поддержки интерактивной учебной деятельности, выполняемой студентами при помощи компьютера [30].

3. Социолингвистика: языковое многообразие. Вариационная лингвистика началась с составления карт диалектов и сборников диалектных выражений в последней трети XIX в. Ее методы были похожи на методы, использовавшиеся в то время исторической лингвистикой, с одной существенной отличительной чертой: корпусы диалектов систематически составлялись по определенным критериям. Вероятно, это можно рассматривать как предвестник все еще продолжающейся дискуссии о том, что включать в корпус.

В настоящее время электронные корпусы часто используются в исследованиях языкового многообразия (например, диалектов, социолектов, регистров). Математические методы (например, мультифакторный анализ, т. е. анализ по многочисленным параметрам) полностью базируются на доступности таких данных.

Современная корпусная лингвистика использует и развивает эти методы. Многие исследования и результаты возможны только с применением больших объемов доступных в электронном виде текстов и современной компьютерной техники. Развитие современных интеллектуальных программных систем, предназначенных для обработки текстов естественного языка, также требует большой экспериментальной лингвистической базы. Спрос на корпусные данные совпал с появлением соответствующих технических возможностей.

1.3. История создания лингвистических корпусов

Лингвисты собрали первые корпусы компьютеризированных текстов в 1960-е гг. Первый компьютеризированный корпус – *Брауновский корпус* (The Brown Corpus[1]) – включает 500 текстов из американских книг, газет, журналов, впервые опубликованных в США в 1961 г. Каждый текст в Брауновском корпусе имеет длину 2000 слов (имеется в виду словоупотреблений – tokens), и все собрание включает 1 млн слов (500 текстов по 2000 слов в каждом). Авторы корпуса У. Френсис (W. Francis) и Г. Кучера (H. Kucera) сопроводили его большим количеством материалов первичной статистической обработки: частотным и ал-

[1] Полное название корпуса – The Brown Standard Corpus of American English. Он был разработан в Брауновском университете (Brown University) в США в 1963 г.

фавитно-частотным словарем, разнообразными статистическими распределениями.

Цель создания Брауновского корпуса – обеспечить системное изучение отдельных жанров письменного английского языка и сравнение жанров. Его появление вызвало всеобщий интерес и оживленные дискуссии. В первую очередь они коснулись принципов отбора текстов и состава потенциально решаемых на таком корпусе задач. С одной стороны, он строился на основе статистических критериев, с другой стороны, статистика применялась в сочетании с волевыми решениями создателей корпуса, базирующимися на профессиональной интуиции. Для достижения максимальной объективности этого сложного процесса требовалось построение максимально формализованных, прозрачных для проверки и контроля процедур.

Позднее европейские исследователи составили корпус текстов, впервые опубликованных в Великобритании в 1961 г., следуя тем же принципам: 15 жанров, 500 текстов по 2000 слов (словоупотреблений). Он включал 1 млн слов британского варианта английского языка, и его назвали *корпусом Ланкастер-Осло-Берген* (The Lancaster-Oslo-Bergen Corpus, по названиям британского и двух норвежских университетов, или кратко LOB).

Итак, два самых ранних больших корпуса – это корпусы письменной речи американского и британского вариантов английского языка. Оба корпуса остаются полезными и сейчас, на них основываются многочисленные исследования английского языка.

За десятилетия, прошедшие с момента создания этих корпусов, компьютеры стали дешевле и гораздо мощнее, кроме того, недорогие и надежные сканеры сделали необязательным набор текстов на компьютере с помощью клавиатуры. Эти достижения облегчили процесс создания корпусов, и последние из них содержат уже миллиарды слов (словоупотреблений).

К 1990 г. уже было зафиксировано более 600 компьютерных корпусов. По годам составления они были распределены примерно следующим образом [44] (табл. 1).

Очевидно, что в последующие годы количество и многообразие создаваемых корпусов шли по нарастающей. Среди современных корпусов *английского* языка (как британского, так и американского варианта) наиболее известны *Британский национальный корпус* (British National Corpus – BNC), *Международный корпус английского языка* (International Corpus of English – ICE), *Лингвистический банк английского языка* (Bank of English), *Корпус современного американского ан-*

глийского (Corpus of Contemporary American English – COCA) и др. В настоящее время корпусы созданы для многих языков мира (см. Приложение 1).

Таблица 1. Количество существующих корпусов в определенные периоды времени

Период	Количество корпусов
До 1965 г.	10
1966–1970	20
1971–1975	30
1976–1980	80
1981-1985	160
1986–1990	320

В первой половине 1990-х гг. корпусная лингвистика окончательно сформировалась как отдельное направление науки о языке. «Корпусная лингвистика достигла зрелости» – так Дж. Свартвик (J. Svartvik) озаглавил в 1992 г. предисловие к материалам первого Нобелевского симпозиума по корпусной лингвистике [60]. Корпусная лингвистика тесно взаимодействует с компьютерной лингвистикой, используя ее достижения и, в свою очередь, обогащая ее.

1.4. Основные характеристики корпусов

1.4.1. Репрезентативность корпусов

Термин «корпус» обычно обозначает собрание текстов конечного фиксированного размера. С течением времени объем и состав корпуса может меняться, однако эти изменения должны либо не менять его структуру, либо менять ее обоснованно. Представительность корпуса, соотношение его отдельных частей (по разным характеристикам) получили название ***репрезентативность***, или ***сбалансировнность***. Объем первых корпусов, как уже говорилось, составлял 1 млн словоупотреблений (Брауновский корпус, корпус Ланкастер-Осло-Берген, Упсальский корпус русского языка). Такой объем не позволял отразить язык во всем его многообразии. В настоящее время считается, что общеязыковой (национальный) корпус должен включать не менее 100 млн словоупотреблений. Национальный корпус представляет данный язык на определенном этапе (или этапах) его существования во всем многообразии жанров, стилей, территориальных и социальных

вариантов и т. п. Можно сказать, что все современные лингвистические исследования и работы по составлению словарей и грамматик так или иначе ориентированы на использование представительных (репрезентативных) корпусов текстов.

Задача авторов корпуса – собрать как можно большее количество текстов, относящихся к тому подмножеству языка, для изучения которого корпус создается. Можно сказать, что корпус – это уменьшенная модель языка. Под ***репрезентативностью*** понимается необходимо-достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов и т. д., т. е. способность отражать все свойства проблемной области [31]. Имеются разные подходы к определению репрезентативности. В частности, можно сказать, что применительно к общеязыковому (национальному) корпусу это понятие невозможно рассчитать и описать строго, однако к этому можно и нужно стремиться, как на этапе проектирования корпуса, так и на этапе его эксплуатации.

Практика показывает, что корпусная лингвистика оперирует как минимум двумя разными типами корпусов текстов:

1) корпусы первого типа – универсальные, отражающие в себе все многообразие речевой деятельности;
2) корпусы второго типа – специфичные, отражающие бытование некоторого языкового или культурного явления в общественной речевой практике, построенные ad hoc (для специальной цели), например, корпус пословиц или корпус политических метафор в газетной речи;
3) корпусы третьего типа – специфичные, создаваемые для решения специальной задачи, например, для отладки систем машинного перевода.

Во всех случаях репрезентативность рассматривается только как статистическая оценка того, все ли свойства проблемной области и решаемой задачи отражены в корпусе текстов. Однако статистические критерии оценки здесь не всегда являются единственными или определяющими, поскольку корпус выступает как некоторый объект, призванный послужить *моделью* внешней по отношению к нему реальности. Именно репрезентативность корпуса определяет достоверность полученных на его материале результатов. Эту проблему также можно рассматривать как проблему адекватного отражения, адаптации или интеграции больших массивов текстов или некоторых иных фрагментов речевой деятельности в существенно меньший по объему корпус текстов.

Речевая действительность чрезвычайно разнообразна, и разнообразие зафиксированных в ней лингвистических явлений просто необозримо. В 1960-е гг. корпусы текстов, относящиеся к *первому типу*, претендовали на то, что они суть универсальные, т. е. отражают статистически корректно всю картину бытования данного языка или некоторый представительный ее фрагмент [51]. Например, Брауновский корпус текстов был создан для отражения письменной речи США 1960-х гг. с удовлетворительной для того времени степенью репрезентативности. Отобранные тексты представляли 15 жанров (регистров), из которых было сделано от 6 до 80 элементарных выборок:

1) пресса: репортаж;
2) пресса: передовица;
3) пресса: обзоры;
4) религиозные тексты;
5) навыки, занятия, хобби;
6) научно-популярная литература;
7) беллетристика, биографии, эссе;
8) разное (правительственные документы, отчеты предприятий, промышленные отчеты, каталоги колледжей);
9) научные сочинения;
10) художественная литература;
11) мистика и детективы;
12) научная проза;
13) приключенческая литература и вестерны;
14) любовные романы;
15) юмористические произведения.

В корпусах *второго типа* критерием репрезентативности будет служить требование максимально полного и объективного представления бытования интересующего его создателей явления. Так, корпус английских пословиц максимально репрезентативно отражает их употребление в речевой практике носителей английского языка определенного времени и географического региона. Наполнение корпусов *третьего типа* определяется спецификой той задачи, для решения которой они создаются. Например, корпус текстов для отладки системы заказа билетов, будет отличаться специфической лексикой и т. п.

Методология конструирования такого объекта, как корпус, должна зависеть от типа корпуса. Эта проблема является актуальной и недостаточно проработанной. Методология построения корпусов первого типа так или иначе основывается на принципе дедукции – реализации

проблемы корректности движения от общего (объективно существующей речевой практики носителей языка) к отражающему это общее частному корпусу текстов. Методология построения корпусов второго и третьего типа должна корректно отражать частные, единичные языковые явления в корпусе текстов, специально созданном для их отражения. Теория и практика показывают, что оба эти подхода, тем не менее, часто применяются в комбинированном виде.

1.4.2. Классификация корпусов по различным основаниям

Несмотря на разнообразие корпусов, можно выделить два основных способа их деления на классы:

1) противопоставление корпусов, относящихся ко всему языку (часто к языку определенного периода), корпусам, относящимся к какому-либо жанру, стилю, языку определенной возрастной или социальной группы, языку писателя или ученого и т. д.;

2) разделение корпусов по типу лингвистической разметки. Несмотря на наличие множества типов разметки, большинство реально существующих корпусов относится к корпусам морфологического либо синтаксического типа (последние в англоязычной литературе называют treebanks). При этом следует подчеркнуть, что корпус с синтаксической разметкой явно или неявно включает в себя и морфологические характеристики лексических единиц.

Существует большое число разных типов корпусов, что определяется многообразием исследовательских и прикладных задач, для решения которых они создаются, и различными основаниями для классификации. В зависимости от поставленных целей и классифицирующих признаков можно выделить различные типы корпусов (табл. 2).

Таблица 2. Классификация корпусов

Признак	Тип корпуса
Цель	многоцелевые специализированные
Тип языковых данных	письменные устные (речевые) смешанные
«Литературность»	литературные диалектные разговорные терминологические смешанные

Жанр	литературные фольклорные драматургические публицистические
Назначение	исследовательские иллюстративные
Динамичность	динамические (мониторные) статические
Разметка	размеченные неразмеченные
Характер разметки	морфологические синтаксические семантические анафорические просодические и т. д.
Доступность	свободно доступные коммерческие закрытые
Объем текстов	полнотекстовые «фрагментнотекстовые»

Естественно, данная классификация весьма условна и неполна.

Итак, *по цели* создания корпусы делятся на многоцелевые и специализированные. Многоцелевые корпусы обычно содержат тексты различных жанров (сюда относятся национальные корпусы), в то время как специализированные корпусы могут ограничиваться одним жанром или группой жанров.

Пример терминологического корпуса – корпус текстов по корпусной лингвистике, позволяющий разрабатывать терминологический словарь непосредственно на живом текстовом материале текстов по корпусной лингвистике [54]. В этом корпусе методология корпусной лингвистики применена к ней самой.

Корпусы текстов могут быть классифицированы *по жанрам* и подразделяться на литературные, фольклорные, драматургические, публицистические и др. Примером публицистического корпуса может служить *Компьютерный корпус текстов русских газет конца XX века* (http://www.philol.msu.ru/~lex/corpus/).

По назначению выделяют исследовательские и иллюстративные корпусы. Исследовательские корпусы создаются с целью изучения различных аспектов функционирования языка. Этот тип корпусов ориен-

тирован на широкий класс лингвистических задач. Неспецифицированность задачи требует корпусов достаточно большого объема. Как правило, такие корпусы текстов содержат от нескольких десятков миллионов до сотен миллионов словоупотреблений.

Иллюстративные корпусы создаются после проведения научного исследования: их цель не столько выявить новые факты, сколько подтвердить и обосновать уже полученные результаты. Они служат для выделения из них лингвистических примеров, подтверждающих те или иные языковые (речевые, текстовые) факты, обнаруженные ранее иными лингвистическими приемами [2].

Типичный пример иллюстративного корпуса представлен в «Путеводителе по дискурсивным словам русского языка» [3], где семантический анализ частиц и выделенные значения сопровождаются обширным текстовым материалом, что позволяет читателю проверить семантические интерпретации, предложенные авторами.

Критерий *динамичность* подразделяет корпусы на динамические и статические. Первоначально корпусы текстов создавались как статические образования, отражающие определенное временно́е состояние языковой системы. Статические корпусы содержат тексты какого-то временно́го промежутка. Типичными представителями этого вида корпусов являются авторские корпусы – коллекции текстов писателей.

Значительная часть чисто лингвистических и не только лингвистических задач требует выявления функционирования языковых явлений на временно́й шкале – например, изменения значения слов, частоты использования тех или иных синтаксических конструкций и т. д. Для отражения процессуального аспекта проблемной области разрабатываются технологии построения и эксплуатации динамических корпусов текстов [2]. Динамические корпусы называют также мониторными или мониторинговыми. Цель их – постоянно наращивать свой объем. В течение заранее фиксированного промежутка времени происходит обновление и/или дополнение множества текстов корпуса.

Неограниченные (постоянно развивающиеся) мониторные корпусы играют огромную роль в строении словарей, поскольку позволяют лексикографам следить за новыми словами, проникающими в язык, или за уже существующими словами, меняющими свое значение, а также за балансом их употребления в соответствии со стилем. Динамические корпусы текстов предназначены для проведения различных диахронических исследований.

Критерий *разметка* делит корпусы на размеченные и неразмеченные. Существуют и другие термины, обозначающие это деление: индексированные и неиндексированные, аннотированные и неаннотированные, тегированные и нетегированные. В размеченном корпусе словам или предложениям присваиваются метки (теги) в соответствии с *характером разметки*: морфологические, синтаксические, семантические, просодические и др.

Важным критерием для пользователей корпуса является его *доступность*. Свободно доступные корпусы позволяют в любое время в режиме online искать по всем текстам корпуса в полном объеме. В ряде случаев свободный доступ может предоставляться к части корпусных данных и не со всеми функциональными возможностями. В работе с коммерческими корпусами нужно покупать право его использования on-line или копию на компакт-диске. Предварительно можно ознакомиться с аннотацией к корпусу или, возможно, даже поработать с корпусом в пробном режиме, но, как правило, не со всеми текстами, а только с небольшим по объему подкорпусом. Закрытые корпусы создаются для специфических целей и не предназначены для публичного использования.

По критерию *объем текстов* выделяют полнотекстовые и так называемые фрагментотекстовые корпусы. Как известно, Брауновский корпус и корпус Ланкастер-Осло-Берген должны были строго соответствовать определенным критериям, одним из которых была длина текста, равная 2000 слов (словоупотреблений). Очевидно, что текстов, строго соответствующих таким критериям, практически нет. Следовательно, эти корпусы являются фрагментотекстовыми. К полнотекстовым корпусам относится большинство современных корпусов, а также корпусы текстов определенного автора.

Полнотекстовыми являются и корпусы специальных коротких текстов, например, *Берлинский корпус перемен* (Berliner Wendekorpus), сформированный с целью создания коллекции личного опыта участия в социальном переломе, известном под названием «Разрушение стены 1989 года», или корпус мерфизмов (так называемых «законов подлости») [5].

Выделяется еще несколько типов корпусов, которые заслуживают особого внимания, поэтому их необходимо рассмотреть более подробно. Критериями для выделения этих разновидностей корпусов являются, соответственно, *параллельность* и *тип языковых данных*.

1.4.3. Особенности корпусов отдельных типов

1.4.3.1. Параллельные корпусы

По критерию *параллельность* корпусы делятся на одноязычные, двуязычные и многоязычные. В одноязычных корпусах противопоставляются диалекты, варианты языка. Например, такие разновидности английского языка, как английский как родной и английский как иностранный, оставались за пределами научного интереса до появления новых технологий, позволивших вовлечь в контрастивный анализ существенно большее количество сопоставляемых произведений речи.

Двуязычные и многоязычные корпусы можно разделить на два основных типа:

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

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

Корпусы обоих типов используются в целях разработки эффективных методов перевода, в том числе, машинного, для составления двуязычных и многоязычных терминологических словарей, а также для сравнительных исследований языков (в области лексикологии, грамматики, стилистики, переводоведения и т. д.).

При подготовке параллельных корпусов первого типа и разработке программ для их обработки возникает проблема выравнивания (alignment) – установление соответствий между фрагментами текста оригинала и текста перевода. Для решения этой задачи используются различные методы автоматического выравнивания текстов – по предложениям, клаузам (грамматическим конструкциям), словосочетаниям и словам. При выравнивании на уровне предложений могут использоваться, как это описано в учебнике А.В. Зубова и И.И. Зубовой [17]: шесть возможных соответствий между предложениями обоих текстов:

1) одно исходное предложение переводится одним предложением;
2) два исходных предложения переводятся одним предложением;
3) одно исходное предложение переводится двумя предложениями;
4) два исходных предложения переводятся двумя предложениями, но внутренние границы этих предложений в тексте оригинала и тексте перевода не совпадают;
5) предложение исходного текста не переводится;

6) предложение в тексте перевода не имеет эквивалента в тексте оригинала.

На практике существуют различные программы выравнивания, которые автоматически сопоставляют тексты на основе совпадения относительных длин предложений, разделения текста на абзацы, анализа знаков препинания, внешнего словаря и других факторов. Чаще всего эти программы используются в человеко-машинном варианте, с постредактированием результатов автоматического выравнивания.

Параллельные корпусы текстов позволяют получить большой объем информации. С их помощью можно:

- строить двуязычные и многоязычные переводные словари;
- создавать и пополнять словари для систем машинного перевода;
- устранять полисемию лексических единиц путем компьютерного анализа контекста многозначного слова, превышающего по длине предложение;
- переводить терминологические и фразеологические единицы текста;
- осуществлять полностью автоматический перевод в рамках новых систем машинного перевода, называемых системами с переводческой памятью, путем накопления в памяти компьютера корпусов исходных текстов и их переводов, выровненных между собой на различных уровнях.

В процессе перевода такая система пытается отыскать переводимое предложение или его фрагмент в массиве исходных параллельных текстов. Если оно найдено в исходном массиве текстов-оригиналов, то система выбирает перевод такого предложения или его части в массиве переведенных текстов [17].

При исследовании параллельных корпусов, в том числе корпусов второго типа, могут успешно применяться инструменты автоматической классификации лексики.

Автоматическая классификация лексики является одной из ключевых процедур автоматического понимания текстов [4]. Она осуществляется в рамках формализации понятийной структуры текста и количественной оценки семантических связей между элементами текста (словами, представленными леммами и словоформами). Сравнительный анализ количественных данных об употреблении слов, о степени их семантической близости помогает устанавливать распределение лексических единиц *разных* языков внутри лексико-семантических и тематических групп. Информация о соотношении элементов кластеров, полученная при параллельной обработке текстов оригинала и перевода в параллельных

корпусах первого типа, имеет высокую ценность в определении адекватности перевода и при проведении контрастивных исследований. Применение модулей автоматической классификации лексики повышает эффективность поиска в параллельных корпусах, позволяет извлекать данные для пополнения и корректировки многоязычных словарей, для проверки качества работы систем машинного перевода и их обучения [25; 7].

Система машинного перевода текста может быть основана на расширенных морфологических союзах между двумя языками с использованием простых правил для выбора подходящих грамматических пар. Например, в параллельном русско-словацком корпусе текстов снятие семантической и морфологической омонимии проводится с применением цепи Маркова первого или второго порядка, которая тренирована на большом одноязычном корпусе. Генетические сходства между лексическими системами русского и словацкого языков можно использовать также для увеличения качества перевода при помощи схемы транслитерации отсутствующих в словаре слов. Система машинного перевода также может учитывать синтаксические сходства между более или менее родственными естественными языками. В частности, это касается таких языков, как чешский и словацкий. Системы переводческой памяти могут быть использованы творчески для большей автоматизации переводческого процесса, не зависящей от конкретных языков.

Параллельные корпусы часто создаются на основе текстов, используемых в многоязычных сообществах, таких как Организация Объединенных Наций, в странах Европейского Союза и в официально двуязычных странах, таких как Канада.

1.4.3.2. Корпусы устной речи

По типу языковых данных корпусы делятся на письменные, устные (речевые) и смешанные. В письменных корпусах устная речь не представлена (Брауновский корпус, LOB), в устных корпусах представлена только устная речь, смешанными обычно бывают национальные корпусы, представляющие бытование языка в определенный период времени (НКРЯ, BNC и др.).

Составители корпуса не всегда могут представить себе все многообразие лингвистических задач, которые могут быть решены с его помощью. Среди них областью особой важности, основной для понимания языка вообще, является исследование устных текстов. Прагматика не была так тщательно исследована в компьютерной лингвистике и корпусных исследованиях, как некоторые другие сферы лингвистики,

поскольку создание репрезентативного корпуса устной речи было сложной задачей. В конце концов возникла необходимость в создании моделей вежливости, смены ролей и других явлений [42].

Первый корпус устной речи *Лондон-Лунд* (The London-Lund Corpus) был разработан в рамках проекта «Обзор употребления английского языка» (The Survey of English Usage). Цель проекта заключалась в том, чтобы по возможности полно зафиксировать особенности грамматической системы английского языка в речи взрослого образованного носителя. Проект разрабатывался с 1959 г. под руководством Р. Квирка (R. Quirk) в Лондонском университетском колледже. Объем корпуса – 1 млн словоупотреблений. Текстами устной речи были записи радиопередач, заседаний официальных структур, а также неформальных бесед. Машинный вариант корпуса создавался в Лундском университете (Швеция) и был готов к использованию в 1979 г.

Именно корпус устной речи Лондон-Лунд был одним из первых машиночитаемых корпусов. Он состоял из 34 текстов, представляющих тайно записанные разговоры, которые были также опубликованы в книге Дж. Свартвика и Р. Квирка «Корпус английского разговора» (1980) [59]. Эта книга была очень полезна в то время, когда компьютерные корпусы не были широко распространены, и было трудно обращаться со сложной транскрипцией устной речи [44]. Хотя некоторой частью информации пришлось пожертвовать при составлении машиночитаемой версии, и те, кого записали, вряд ли могут считаться среднестатистическими представителями лиц, говорящих на английском языке, корпус Лондон-Лунд очень помог в изучении речи. Из-за сложностей составления корпусов устной речи этот корпус долго оставался самым важным источником для компьютерного исследования разговорного английского.

Появление корпуса Лондон-Лунд привело к множеству исследований по лексике, грамматике, просодии речи и особенно по структуре и функционированию дискурса. Так, были исследованы использование слов actually, really, you know, you see, I mean, well, вопросы и ответы в английском разговоре, использование пассива, просодических моделей английского разговора и т. д. Устный и письменный английский изучались в сопоставительных исследованиях на базе корпусов Лондон-Лунд и Ланкастер-Осло-Берген; в частности, изучались модальность, связи в сложных предложениях, отрицание.

В настоящее время большой интерес корпусных лингвистов привлекают способы передачи эмоций в устной речи, выражение удивле-

ния и т. д. Примером корпуса, позволяющего проводить подобные исследования, является мультимедийный подкорпус в составе НКРЯ.

Отсутствие баланса в доступности устного и письменного материала в машиночитаемом формате продлится еще очень долго. В силу различных причин построение корпусов устной речи продвигается намного медленнее, чем построение корпусов письменной речи. В первую очередь устную речь нужно как-то зафиксировать – например, с помощью магнитной ленты, цифровой записи или видеокассеты. Затем ее нужно записать буквами, что является утомительной и дорогой работой, качество которой зависит в большой степени от качества записи и степени шума внешней среды в естественных условиях.

Главная сложность создания фонетических лингвистических ресурсов связана с необходимостью транскрибирования устной речи. При этом возникают следующие проблемы:

1) выбор алгоритма для транскрибирования;
2) учет индивидуальных особенностей произношения;
3) учет всего устного текста или его фрагментов;
4) учет диалектных вариантов произношения слов;
5) учет ударений в словах;
6) учет просодических признаков произносимых фраз;
7) маркирование слов, которые при прослушивании не распознавались;
8) маркирование в записи для фонетического корпуса паралингвистических явлений, сопутствующих речи (паузы, смех, бормотание, кашель, и т. д.).

В настоящее время общепринято, что для создания машиночитаемых фонетических корпусов используется *транскрипция на основе орфографического представления звуков* речи с дополнительными знаками, передающими (при необходимости) просодические, паралингвистические и другие особенности произношения.

Несмотря на трудности создания в мире уже существует много достаточно представительных фонетических корпусов. Так, в 70-х гг. XX в. в США Х. Далем и его коллегами был создан «*Корпус устной речи американского варианта английского языка*», который включал 1 млн словоупотреблений, взятых из записей психоаналитических сеансов. С каждой из 15 кассет, имевшихся в распоряжении составителей корпуса, было случайным образом отобрано 225 записей сеансов. Они содержали речь 8 женщин и 21 мужчины из 9 городов США. Отобранные записи были затранскрибированы на основе стандартной английской орфографии. Диалектные варианты произношения не учитыва-

лись. Нераспознанные слова при записи обозначались буквой Z. Ударения и другие просодические характеристики речи также не учитывались. В то же время при орфографической записи устной речи в качестве специальных комментариев отмечались паузы, смех, вздох, кашель и другие паралингвистические явления [17].

Один из членов команды, создававшей Британский национальный корпус, Л. Бёрнард (L. Burnard), утверждал, что стоимость отбора 10 млн слов из устных источников во время создания корпуса (1990-е гг.) равнялась стоимости отбора 50 млн слов из письменных источников [26]. Данные издержки напрямую связаны еще и со строго соблюдаемым в западном мире авторским правом, в связи с чем нельзя провести полноценный анализ устных текстов и опубликовать его результаты без получения согласия их автора, а это не всегда возможно по объективным причинам.

По состоянию на март 2013 г. устный подкорпус Национального корпуса русского языка насчитывает 10,3 млн словоупотреблений. Устный компонент корпуса подразделяется на следующие типы: публичная речь – 5,5 млн словоупотреблений (53 %), непубличная – 1,2 млн (11 %), речь кино – 3,7 млн (36 %), авторское чтение 24 тыс. словоупотреблений, художественное чтение – 2,5 тыс. театральная речь – 4,3 тыс. Мультимедийный корпус объемом 3,5 млн словоупотреблений включает прежде всего речь кино (3,4 млн), а также публичную речь (31,5 тыс.), непубличную речь (12,6 тыс.) и др.

Одной из наиболее важных проблем при составлении национальных корпусов текстов является их недостаточное наполнение устными текстами, особенно относящимися к непубличной речи – телефонным разговорам, неформальным беседам и т. д. Примером такого речевого корпуса является корпус «Один речевой день» (ОРД), разрабатываемый в Санкт-Петербургском университете [38]. Подробнее см. п. 3.2.2.2.5.

Вопросы для самоконтроля

1) Дайте определения терминов: *корпус*, *корпусная лингвистика*, *разметка*, *репрезентативность*.
2) Перечислите типы корпусов.
3) Какой корпус текстов был первым? Укажите его основные характеристики.
4) Какие два типа корпусов можно выделить по критерию репрезентативности? В чем их отличие?
5) Каково основное назначение параллельных корпусов текстов?
6) В чем заключается сложность создания корпусов устной речи?

Часть 2
СОЗДАНИЕ КОРПУСОВ

2.1. Предварительные работы по созданию корпуса

2.1.1. Проектирование и технологический процесс создания

Проект любого корпуса должен предусматривать этапы его создания и пути его дальнейшего развития. Понятие корпуса является продолжением традиционных картотек, с которыми всегда работали лингвисты. В XX в. эти картотеки стали компьютерными и общедоступными. Значительную роль в становлении корпусного подхода сыграла сеть Интернет, в процессе развития которой стали доступны большие объемы текстового материала, пригодного для проведения различных лингвистических исследований.

При проектировании корпуса должен быть решен ряд вопросов, касающихся наполнения и структуры корпуса. Прежде всего это традиционный вопрос о репрезентативности и сбалансированности языкового материала (см. п. 1.4.1), который кладется в основу словарей и грамматик, создаваемых на базе корпуса. Особенно остро этот вопрос встает при формировании национальных корпусов. Репрезентативность корпуса должна обеспечиваться как достаточным объемом текстового материала, так и его разнообразием.

Не менее важна и проблема хронологии. Что следует понимать под корпусом *современного* языка? Представляется, что хронологические рамки корпуса должны быть разными для разных жанров.

Корпус создается для широкого круга пользователей и для решения разнообразных задач, в том числе и достаточно «экзотических», например, для исследования русскоязычных текстов, использующих иноязычную графику. Что из исходных текстов остается в корпусе, а что «вычищается»? Очевидно, например, что картинки не относятся к языковому материалу и могут быть удалены. Сложнее обстоит дело с таблицами и, тем более, с цитатами, прямой речью, иноязычными вкраплениями, единицами измерения.

Все эти вопросы должны быть поставлены на этапе проектирования. Решаться же они, по крайней мере, некоторые из них, могут постепенно в процессе создания и опытной эксплуатации корпуса. Для этого с самого начала эксплуатации следует предусмотреть обратную связь с пользователями.

Технологический процесс создания корпуса можно представить в виде следующих шагов или этапов.

1. Обеспечение поступления текстов в соответствии с перечнем источников.

2. Преобразование в машиночитаемую форму. Тексты в электронном виде для создания корпусов могут быть получены самыми разными способами – ручной ввод, сканирование, авторские копии, дары и обмен, Интернет, оригинал-макеты, предоставляемые составителям корпусов, и др.

3. Анализ и предварительная обработка текстов. На этом этапе все тексты, полученные из разных источников, проходят филологическую выверку и корректировку. Подготовка «технологического» описания сопровождается здесь с библиографическим описанием текста.

4. Конвертирование. Некоторые тексты проходят также через один или несколько этапов предварительной машинной обработки, в ходе которых осуществляется перекодировка (если требуется), а также удаление или преобразование нетекстовых элементов (рисунки, таблицы), удаление из текста переносов, «жестких концов строк» (тексты из MS-DOS), обеспечение единообразного написания тире и т. д.

5. Графематический анализ (токенизация). Он предполагает проведение следующих операций: разделение входного текста на элементы (слова, разделители и т. д.), удаление нетекстовых элементов, выделение и оформление нестандартных (нелексических) элементов, обработка специальных текстовых элементов (имен, написанных инициалами, иностранных лексем, записанных латиницей, названий рисунков, примечаний, страниц форзаца, зачеркиваний, титульных листов, списков литературы и т. д.). Как правило, эти операции выполняются в автоматическом режиме. Обычно на этом же этапе осуществляется сегментирование текста на его структурные составляющие.

6. Разметка текста. Она заключается в приписывании текстам и их компонентам дополнительной информации (метаданных). Метаданные можно поделить на 3 типа: экстралингвистические, относящиеся ко всему тексту; данные о структуре текста; лингвистические метаданные, описывающие элементы текста. Метаописание текстов корпуса включает как содержательные элементы данных (библиографические дан-

ные, признаки, характеризующие жанровые и стилевые особенности текста, сведения об авторе), так и формальные (имя файла, параметры кодирования, версия языка разметки, исполнители этапов работ). Экстралингвистические метаданные являются результатом интеллектуальной обработки текстов и вводятся вручную. Структурная разметка документа (выделение абзацев, предложений, слов) и собственно лингвистическая разметка обычно осуществляются автоматически.

7. Корректировка результатов автоматической разметки: исправление ошибок и снятие неоднозначности (вручную или полуавтоматически).

8. Конвертирование размеченных текстов в структуру специализированной лингвистической информационно-поисковой системы (corpus manager), обеспечивающей быстрый многоаспектный поиск и статистическую обработку (заключительный этап).

9. Обеспечение доступа к корпусу. Корпус может быть доступен в пределах дисплейного класса, может распространяться на компакт-диске и может быть доступен в режиме глобальной сети. Различным категориям пользователей могут предоставляться разные права и разные возможности.

10. Создание документационного обеспечения, в котором описываются различные аспекты создания и использования корпуса, в частности, приводятся сведения о разметке, позволяющие искать по метаданным, язык запросов корпус-менеджера и т. д.

Конечно, в каждом конкретном случае состав и количество процедур могут отличаться от вышеперечисленных, и реальная технология может оказаться гораздо сложнее. Рассмотрим некоторые этапы более подробно.

2.1.2. Отбор источников. Критерии отбора

Важной особенностью корпуса текстов является то, что это не просто множество случайным образом объединенных текстов того или иного языка. При его создании должен быть разрешен целый ряд проблем. Основными из них являются следующие:

1. Что является основной единицей корпуса?
2. Каким должен быть объем корпуса текстов (сколько единиц он должен содержать)?
3. Какие письменные текстовые источники должны быть представлены в корпусе текстов и в каком количестве?
4. Из какой исходной языковой области должны быть выбраны тексты, включаемые в состав корпуса?

Первые ответы на эти вопросы в отечественной лингвистике были даны в многочисленных исследованиях профессора Р.Г. Пиотровского и его коллег в 1965–1980 гг., они касались отбора текстов для составления частотных словарей и проведения лингвостатистических исследований. Именно тогда были впервые использованы различные статистические приемы для оценки генеральной совокупности выборки, объема выборки, порции выборки (элементарной выборки) и т. д. [17]. Схожие проблемы решались при создании «Частотного словаря русского языка» под руководством Л.Н. Засориной – см. предисловие к этому словарю [13].

Основной единицей корпуса текстов являются *словоупотребления* (или токен, англ. tokens, часто их называют словами, words), но также корпусная лингвистика оперирует понятиями *словоформа* (type), *основная форма, лемма* (lemma), *оборот, устойчивое словосочетание* (multiword unit), *предложение* (sentence). Объем создаваемого корпуса текстов в принятых единицах зависит от целей создания. Он может быть небольшим при изучении частоты употребления букв, буквосочетаний, звуков, звукосочетаний. Гораздо бóльшим он должен быть при изучении лексики, морфологических явлений и при изучении синтаксических или стилистических особенностей текстов. Проблемными являются также следующие вопросы:

1. Тексты каких функциональных жанров включать в корпус текстов (художественную прозу, драму, стихи, научные тексты, газеты, журналы, технические описания и т. п.)?
2. Тексты каких временны*х* промежутков включать в национальный (общеязыковой) корпус текстов (современные, 10-летней давности, 50-летней давности, старше и т. д.)?
3. Включать ли тексты только литературного языка или также другие типы источников? И что считать литературным языком?

При ответе на эти вопросы разработчики корпуса текстов, помимо своих идей, используют консультации специалистов по языкознанию и лингвостатистике или метод анкетирования. Исходя из своего опыта исследований, специалисты определяют общий объем корпуса текстов, время издания текстов, число текстов и размер элементарной выборки, жанры отбираемых текстов и их количество, число элементарных выборок из каждого жанра. Метод анкет в сочетании с опытом специалистов был использован при создании корпуса текстов «Базовый корпус американского наследия» (The American Heritage Intermediate Corpus). Специалисты определили его объем в 5 млн слов (словоупотреблений) и рекомендовали включить в него лексику из 22 разделов (жанров)

детской и юношеской литературы на английском языке. В 221 школу США были разосланы анкеты с просьбой указать, какие тексты желательно включить в корпус. После изучения анкет был составлен список из 19 тыс. названий книг. Из этого множества было отобрано 1045 текстов. На их основе было составлено 10 тыс. элементарных выборок по 500 словоупотреблений каждая [17].

2.1.3. Основные процедуры обработки естественного языка: токенизация, лемматизация, стемминг, парсинг

В процессе создания корпуса естественно использование специальных процедур и программ. Например, ***токенизация*** **(**или ***графематический анализ*****)**, т. е. разбиение потока символов в текстах на естественном языке на отдельные значимые единицы (токены, словоформы), является необходимым условием для дальнейшей обработки текстов. Если бы языки обладали совершенной пунктуацией, токенизация не представляла бы сложности – даже самая простая программа могла бы разделить текст на слова, руководствуясь пробелами и знаками препинания. Но в действительности языки подобной делимитацией не обладают, что усложняет задачу токенизации. Например, в английском языке встречаются случаи, которые не могут быть однозначно токенизированы. Ср.: строка chap. может являться сокращенной формой слова chapter или словом chap, которое расположено в конце предложения. Строку Jan. можно рассматривать как сокращенную форму слова January либо как имя собственное, расположенное в конце предложения. В первом случае точка должна быть отнесена к тому же токену, что и слово, а во втором случае она должна быть выделена в отдельный тэг.

Следует заметить, что на этапе токенизации должно быть решено множество мелких, но зачастую важных проблем, и что многие приложения, обрабатывающие текст, нередко игнорируют трудные случаи (например, обработка дефисов, учет аббревиатур и сложных слов, написание в разрядку, многословные лексические единицы и т. п.) либо обрабатывают их с помощью специальных алгоритмов в процедуре поиска.

Другая специфическая задача, относящаяся к *морфологической разметке* – это ***лемматизация***, т. е. процедура образования первоначальной формы слова для словоформ текста. Во многих языках слово может встречаться в нескольких формах с различными флексиями. Например, английский глагол walk может быть представлен следую-

щими формами: walk, walked, walks, walking. Базовая форма, walk, зафиксированная в словаре, называется ***леммой*** слова.

Процесс, несколько отличный от лемматизации, называется ***стеммингом***, он состоит в нахождении стема (основы) слова. Разница заключается в том, что стеммер не порождает нормальную форму, а лишь пытается отсечь изменяемую часть слова, оставив для последующей обработки основу.

Ниже приведены примеры стемминга и лемматизации. Дано следующее предложение:

[The] [quick] [brown] [fox] [jumped] [over] [the] [lazy] [dogs].

Один из наиболее популярных стеммеров, SnowballAnalyzer, выдает следующие стемы:

[quick] [brown] [fox] [jump] [over] [lazy] [dog].

Леммы слов данного предложения будут следующими:

[the] [quick] [brown] [fox] [jump] [over] [the] [lazy] [dog].

Стеммеры обычно более просты для реализации и быстрее обрабатывают данные, но при этом показывают более низкую точность работы. Например, токену better соответствует лемма good, но при стемминге, скорее всего, будет порожден стем bett. Для ряда приложений это может не иметь решающего значения, но в корпусной лингвистике, как правило, имеет место лемматизация. Стеммеры неплохо работают на текстах агглютинативных языков, в то время как для флективных языков, в частности славянских, их результаты уже не так хороши.

Парсинг – это процесс сопоставления линейной последовательности лексем (слов, токенов) языка с его формальной грамматикой. Результатом обычно является дерево зависимостей (синтаксическое дерево) или дерево составляющих. Построение автоматических синтаксических анализаторов (парсеров) для больших корпусов является одной из самых важных областей компьютерной лингвистики.

Наряду с разными статистическими программами, которые «тренируются» на размеченных вручную синтаксических корпусах, многие синтаксические анализаторы используют подходы, основанные на контекстных (лингвистических) правилах, которые моделируют специфические лингвистические теории и грамматические и синтагматические правила построения текстов. Разработка синтаксических анализаторов тесно переплетается с развитием этих теорий. Поскольку большинство предложений неоднозначны в любой теории, на основе правил (или перечня ограничений) должна быть разработана стратегия снятия неоднозначности. Многие стратегии снятия неоднозначности

полагаются на количественные данные – частоту данной структуры в данном корпусе, ограничения на выборку для данных лексических единиц, которые были получены или выделены из корпусных данных, и т. д.

Необходимо рассматривать два условия при обсуждении процедур обработки текстов корпусов:

1. Каждый шаг обработки текста заставляет составителя корпуса принимать лингвистические решения, которые влияют на последующие шаги и на оценку корпуса. Конечный пользователь должен быть в курсе этих решений, чтобы найти то, что он ищет. Например, тот, кто делит тексты на составные элементы (вручную или автоматически), должен решить, являются ли словосочетания New York и Baden Baden, в виду и потому что одним словом или двумя. Подобным образом должны быть приняты решения, как обрабатывать такие явления, как аналитические временные формы в русском языке, немецкие глаголы с отделяемыми приставками и т. п.

2. Конечного пользователя нужно поставить в известность о том, какая работа была проделана на стадии предварительной обработки и о возможных погрешностях, поскольку любые ошибки в кодировке, особенно системные, могут повлиять на результаты, полученные пользователями корпуса [42].

2.2. Разметка. Средства разметки корпусов

2.2.1. Понятие разметки

Среди специальных программ для обработки естественного языка особое место занимают программы автоматической разметки. Разметка корпусов (tagging, annotation) представляет собой трудоемкую операцию, особенно при огромных размерах современных корпусов. Если для некоторых видов разметки, в частности, анафорической, просодической, создание автоматических систем пока представляется довольно сложным и основная часть работы проводится вручную, то для морфологического и синтаксического анализа существуют различные программные средства, которые принято называть соответственно *тэггеры* (taggers) и *парсеры* (parsers).

В результате работы программ автоматической морфологической разметки каждой лексической единице приписываются грамматические характеристики, включая часть речи, лемму и набор граммем

(например, род, число, падеж, одушевленность/неодушевленность, переходность/непереходность и т. д.).

В результате работы программ автоматической синтаксической разметки фиксируются синтаксические связи между словами и словосочетаниями, а синтаксическим единицам приписываются соответствующие характеристики (тип предложения, синтаксическая функция словосочетания и т. п.).

Однако автоматический анализ естественного языка небезошибочен и неоднозначен – он, как правило, дает несколько вариантов анализа для одной и той же языковой единицы (слова, словосочетания, предложения). В этом случае говорят о грамматической омонимии. Снятие неоднозначности (морфологической, синтаксической) в целом является одной из важнейших и сложнейших задач компьютерной лингвистики. При создании корпусов для снятия неоднозначности используются автоматический и ручной[1] способы обработки.

Корпусы нового поколения включают сотни миллионов слов, поэтому требуются системы, которые бы минимизировали вмешательство человека. Автоматическое разрешение морфологической или синтаксической неоднозначности, как правило, основывается на учете контекста и использовании информации более высокого уровня (синтаксического, семантического) с применением статистических методов.

Покажем подходы к снятию неоднозначности на примере английского слова deal. Как словоформа оно может быть и существительным, и глаголом. Предположим, что корпус содержал фразу a good deal of trouble и что автоматическое совмещение со словарем уже позволило пометить good как прилагательное. При выборе части речи для deal программа смотрит, предшествует ли данному слову прилагательное, и если да, то для него значение части речи «существительное», поскольку в английском языке прилагательные обычно предшествуют существительным. Так, deal в a good deal of trouble должно быть помечено как существительное. Если начинать разметку, размечая только слова, принадлежащие исключительно одной категории, а затем использовать эту информацию для того, чтобы прояснить неоднозначные случаи, многие сложные проблемы смогут быть решены. В обычной практике случается так, что сначала слова снабжаются пометами всех частей

[1] Здесь и далее под ручной обработкой понимается дополнительный интеллектуальный анализ с привлечением других методов анализа.

речи, к которым они могут относиться, а затем категории примыкающих слов используются для определения категории слов, у которых есть несколько помет.

Итак, ***разметка*** заключается в приписывании текстам и их компонентам специальных тэгов: собственно *лингвистических*, описывающих лексические, грамматические и прочие характеристики элементов текста, и внешних, *экстралингвистических* (сведения об авторе и сведения о тексте: автор, название, год и место издания, жанр, тематика).

2.2.2. Лингвистическая разметка

Среди лингвистических типов разметки выделяются: морфологическая, синтаксическая, семантическая, анафорическая, просодическая, дискурсная и др. Все они осуществляются в соответствии со следующими принципами:

1) теоретически нейтральная (традиционная) схема разметки;
2) общепринятая система лингвистических понятий;
3) известная для пользователя схема анализа;
4) мотивированность введения параметров;
5) следование международным стандартам.

Морфологическая разметка. В иностранной терминологии употребляется термин part-of-speech tagging (POS-tagging), дословно – частеречная разметка. В действительности морфологические метки включают не только признак части речи, но и признаки грамматических категорий, свойственных данной части речи. Это основной тип разметки: во-первых, большинство крупных корпусов являются как раз морфологически размеченными корпусами, во-вторых, морфологический анализ рассматривается как основа для дальнейших форм анализа – синтаксического и семантического, и, в-третьих, успехи в компьютерной морфологии позволяют автоматически с большой степенью правильности размечать корпусы больших размеров.

Данные о разметке представляются в том или ином структурированном виде и включают: лемму, признак части речи, признаки грамматических категорий. В 1980 г. появилась размеченная версия Брауновского корпуса, в которой была проведена лемматизация словоформ, маркировка их поверхностно-синтаксических функций и т. д.

Морфологическая разметка фрагмента Брауновского корпуса «The jury further said in term-end presentments that the City Executive Committee, which had over-all charge of the election, "deserves the praise and

thanks of the City of Atlanta" for the manner in which the election was conducted.» выглядит следующим образом:

the_AT jury_NN further_RB said_VBD in_IN term-end_NN presentments_NNS that_CS the_AT *city_NP *executive_NP *committee_NP ,_, which_WDT had_HVD over-all_JJ charge_NN of_IN the_AT election_NN ,_, deserves_VBZ the_AT praise_NN and_CC thanks_NNS of_IN the_AT *city_NP of_NP *atlanta_NP for_IN the_AT manner_NN in_IN which_WDT the_AT election_NN was_BEDZ conducted_VBN |

Следом за каждой словоформой через знак подчерк записывается код ее морфологических характеристик (* означает, что следующая за ней буква прописная).

Приведем пример морфологической разметки фрагмента текста на русском языке «*Звонили к вечерне. Торжественный гул колоколов…*» в XML-формате на основе разметчика сервиса АОТ (рис. 1).

```
<?xml version="1.0" encoding="windows-1251" ?> <text> <p>
<s>
<w>Звонили<ana lemma="ЗВОНИТЬ" pos="Г" gram="мн,нс,нп,дст,прш,"/></w>
<w>к<ana lemma="К" pos="ПРЕДЛ" gram="" /></w>
<w>вечерне
<ana lemma="ВЕЧЕРНЯ" pos="С" gram="жр,ед,дт,пр,но," />
<ana lemma="ВЕЧЕРНИЙ" pos="П" gram="ср,ед,кр," /></w>
<pun>.</pun> </s>
<s><w>Торжественный<ana lemma="ТОРЖЕСТВЕННЫЙ"
pos="П" gram="мр,ед,им,вн," /></w>
<w>гул<ana lemma="ГУЛ" pos="С" gram="мр,ед,им,вн,но," /></w>
<w>колоколов
<ana lemma="КОЛОКОЛ" pos="С" gram="мр,мн,рд,но," />
<ana lemma="КОЛОКОЛОВ" pos="С" gram="мр,фам,ед,им,од," /></w>
……..……..
<pun>.</pun> </s></p></text>
```

Рис. 1. Пример морфологической разметки текста на русском языке.

В представленной записи использованы тэги <text> – текст, <p> – абзац, <s> – предложение, <w> – словоупотребление, <pun> – знак пунктуации. Тэг <w> содержит вложенный тэг <ana> с атрибутами <lemma> – лемма, <pos> – часть речи, <gram> – набор граммем.

Синтаксическая разметка. Синтаксическая разметка является результатом парсинга, выполняемого на основе данных морфологического анализа. Этот вид разметки зависит от принятой формальной син-

таксической модели и описывает синтаксические связи между лексическими единицами и/или различные синтаксические конструкции (например, придаточное предложение, именное сказуемое и т. д.).

В отличие от морфологии, способы представления синтаксической структуры и синтаксических отношений не столь унифицированы. Наблюдается разнообразие синтаксических теорий и формализмов:

- грамматика зависимостей;
- грамматика непосредственно составляющих;
- грамматика структурных схем;
- традиционные синтаксические учения о членах предложения;
- функциональная грамматика;
- семантический синтаксис и др.

Синтаксический анализ для русского языка чаще всего представлен структурой зависимостей. На рис. 2 представлен пример визуализации дерева зависимостей предложения *«Потому что из муки высших сортов его не пекут»*

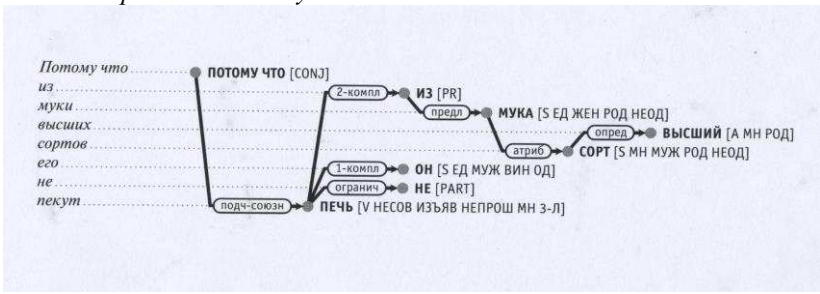


Рис. 2. Пример синтаксического разбора
(грамматика зависимостей, система ЭТАП-3).

Семантическая разметка. Семантические теги чаще всего обозначают семантические категории, к которым относится данное слово или словосочетание, и более узкие подкатегории, специфицирующие его значение. Семантическая разметка корпусов предусматривает спецификацию значения слов, разрешение омонимии и синонимии, категоризацию слов (разряды), выделение тематических классов, признаков каузативности, оценочных и деривационных характеристик и т. д.

Один из вариантов семантической разметки предлагает НКРЯ. В этом корпусе каждой словоформе приписываются пометы трех типов:

1) разряд (имя собственное, возвратное местоимение и т. д.);

2) лексико-семантические характеристики (тематический класс лексемы, признаки каузативности, оценки и т. д.);

3) деривационные характеристики («диминутив», «отадъективное наречие» и т. д.).

Собственно лексико-семантические тэги сгруппированы по следующим полям:

• таксономия (тематический класс лексемы) – для имен существительных, прилагательных, глаголов и наречий;

• мереология (указание на отношения «часть – целое», «элемент – множество») – для предметных и непредметных имен;

• топология (топологический статус обозначаемого объекта) – для предметных имен;

• каузация – для глаголов;

• служебный статус – для глаголов;

• оценка – для предметных и непредметных имен, прилагательных и наречий.

Словообразовательные характеристики включают несколько типов:

• морфо-семантические словообразовательные признаки (например, «каритив», «семельфактив»);

• разряд производящего слова (например, отглагольное существительное или отадъективное наречие);

• лексико-семантический (таксономический) тип производящего слова (например, наречие, образованное от прилагательного размера);

• морфологический тип словообразования (субстантивация, сложное слово) (более подробно о семантической разметке в НКРЯ см. http://ruscorpora.ru, раздел «Семантика»).

Существуют и другие типы разметки, в частности:

• анафорическая разметка. Она фиксирует референтные связи, например, местоименные;

• просодическая разметка. В просодических корпусах применяются тэги, обозначающие ударение и интонацию. В корпусах устной разговорной речи просодическая разметка часто сопровождается так называемой *дискурсной* разметкой, которая служит для обозначения пауз, повторов, оговорок и т. д.

2.2.3. Экстралингвистическая разметка

Экстралингвистическая разметка (метаразметка) включает в себя «внешнюю», «интеллектуальную» разметку (библиографические характеристики, типологические тематические социологические характеристики), которая может дополняться данными технологической раз-

метки (кодировка, даты обработки, исполнители, источник электронной версии).

Набор метаданных во многом определяет дополнительные возможности поиска, предоставляемые корпусами исследователям. При выборе этих данных необходимо руководствоваться целями исследования и потребностями лингвистов, а также возможностями по внесению в текст тех или иных дополнительных признаков.

Метаразметка нужна, во-первых, для выявления взаимосвязи языка и условий его существования; во-вторых, для отбора и изучения отдельных подмножеств языка.

Набор признаков для метаданных чаще всего основывается на рекомендациях проекта TEI (Text Encoding Initiative). Выделяют два класса факторов, влияющих на язык текстов:

- внешние, внеязыковые факторы (**E** – external);
- внутренние факторы (**I** – internal).

Дж. Синклер выделяет три группы **E**-*факторов*:

- E1 (origin) – факторы, относящиеся к созданию текста автором;
- E2 (state) – факторы, относящиеся к внешним признакам текста (включая устную или письменную речь);
- E3 (aims) – факторы, относящиеся к причинам создания текста и его влиянию на аудиторию, и две группы **I**-*факторов*:
- I1 (topic) – предметная область текста;
- I2 (style) – стилистические особенности (стиль, жанр) [57].

В НКРЯ, например, используется следующий набор метаданных.

Первый блок:

1) *автор текста*: имя, пол, дата рождения (или примерный возраст);
2) *название текста*;
3) *время и место создания текста* (может указываться точно или приблизительно);
4) *объем текста*: для художественных произведений приня то, что обычная длина рассказа – менее 5 тыс. слов; обычная длина повести – от 5 до 15 тыс. слов; обычная длина романа – более 15 тыс. слов.

Второй блок: параметры метаописания трех основных *массивов* текстов корпуса – художественных текстов; нехудожественных текстов; драматургических произведений. Например, для художественных текстов в НКРЯ указывается:

1) жанр текста: нежанровая проза, автобиографическая проза, детектив, детская литература, историческая проза, криминальная литература, приключения, фантастика, юмор и сатира;

2) тип текста: автобиографическая проза, анекдот, ассоциативная проза, боевик, детектив, очерк, литературное письмо, повесть, притча, пьеса, рассказ, роман, сказка, триллер, эпопея, эссе и др.;

3) хронотоп текста: приблизительное указание на место и время описываемых в тексте событий [27].

При указании на хронотоп текста в НКРЯ предлагаются следующие опции: древний Восток; Россия XVII век; Россия XIX век; Россия/СССР: советский период в целом; Россия, советский период – Германия 1920–1940-е годы; Россия/СССР – Европа 1960–1980-е годы; Россия/СССР: перестройка; Россия/СССР: советский и постсоветский период; Америка: современная жизнь; Израиль: современная жизнь; Средняя Азия: современная жизнь; ирреальный мир и др. Также может встретиться тэг «хронотоп не определен».

Служебная, или «имплицитная», метаразметка в НКРЯ включает:

1) «текст-стиль», при этом выделяются академический, научно-популярный, официально-деловой, нейтральный, сниженный, сниженный с элементами грубого просторечия и жаргона, индивидуально-авторский, диалектный и пр. (всего 21);

2) аудитория – возраст;

3) аудитория – уровень образования;

4) аудитория – размер.

Более подробно см.: http://ruscorpora.ru/corpora-parameter.html

2.2.4. Стандартизация в корпусной лингвистике

Корпусы, как правило, предназначены для неоднократного применения многими пользователями, поэтому их разметка и их лингвистическое обеспечение должны быть определенным образом унифицированы. Стандарты в отношении корпусов обычно затрагивают совместимость типов разметки. Их называют иногда «стандартами кодирования». Также важным является вопрос, связанный со сравнимостью разных корпусов, в том числе, оценками по поводу их пригодности к различным заданиям. Их называют «стандартами оценки».

Наибольшую сложность представляет стандартизация транскрибирования устной речи и исторических корпусов. Хотя в области графической фиксации устной речи даже при отсутствии единого и обяза-

тельного для всех стандарта достигнут некоторый прогресс (связанный прежде всего с наличием прецедентов), то в описании невербальной составляющей естественноязыковой коммуникации стандарты до сих пор не выработаны, что затрудняет дальнейшее продвижение в этом направлении [2].

Стандартизация в отношении корпусов, совместимость типов данных важны и с точки зрения сравнимости разных корпусов. Причем корпусы могут подвергаться как количественной, так и качественной оценке. Количественные данные о корпусах позволяют судить об их объеме, о наполнении корпуса по различным критериям, о лингвостатистических параметрах корпуса или подкорпусов. Под качественной оценкой понимается оценка и сравнение корпусов на основе анализа выдаваемых результатов.

Вопросы пригодности корпусов к различным лингвистическим заданиям также требуют своих «стандартов оценки». Единые форматы представления данных позволяют во многих случаях использовать единое программное обеспечение и обмениваться корпусными данными.

Можно говорить, с одной стороны, о стандартизации форматов представления данных с точки зрения их наполнения, с другой стороны, с точки зрения их структуры.

Параметры разметки корпусов и их значения должны быть достаточно «естественными», т. е. должны соответствовать общепринятым научным классификациям. Лингвистическое и программное обеспечение корпус-менеджеров должно поддерживать обработку типовых запросов и решение типовых задач.

2.2.4.1. Международные стандарты корпусной лингвистики

В настоящее время на основе международного опыта выработались де-факто стандарты представления метаданных, как лингвистических, так и экстралингвистических, базирующиеся на описаниях - текстов и корпусов в рамках проектов Text Encoding Initiative (TEI), ISLE Project (International Standards for Language Engineering) и на рекомендациях EAGLES (Expert Advisory Group on Language Engineering Standards). Среди них в первую очередь следует назвать CDIF (Corpus Document Interchange Format, www.natcorp.ox.ac.uk/archive/ vault/ tgcw30.pdf), CES (Corpus Encoding Standard, http://www.cs.vassar.edu/ CES/CES1.html#Contents), XCES (Corpus Encoding Standard for XML, http://www.xces.org/).

Эти и другие стандарты в настоящее время «собираются» и обобщаются под эгидой комитета Международной организации по стандар-

тизации ISO/TC 37. Многие из них напрямую относятся к корпусной лингвистике, как-то: ISO 24614-1:2010. Пословная сегментация письменных текстов. Часть 1. Основные концепции и общие принципы; ISO 24610-1:2006. Структуры элементов. Часть 1. Представление структуры элементов данных; ISO 24610-2:2011. Структуры элементов. Часть 2. Описание системы элементов данных; ISO/DIS 24611. Морфосинтаксическая разметка; ISO 24613:2008. Схема лексической разметки; ISO 24615:2010. Система синтаксического аннотирования (SynAF) и др. Эти стандарты под общим названием «Управление лингвистическими ресурсами» описывают:

- принципы и методы стандартизации терминологии,
- разработку терминологических стандартов,
- терминологические словари,
- создание языковых ресурсов,
- компьютерную лексикографию,
- терминологическую документацию,
- кодирование в области терминологии и лингвистических ресурсов,
- использование терминологии и других языковых ресурсов в языковой инженерии и управлении контентом.

2.2.4.2. Формат TEI

Наиболее проработанными являются рекомендации проекта Text Encoding Initiative (TEI). Начало проекта по созданию системы кодирования текстов связано с семинаром в Вассарском колледже в 1987 г., на котором присутствовали представители текстовых архивов, научных обществ и исследовательских центров. Целью встречи было обсуждение возможности создания стандартной схемы кодирования текстовых документов. Собственно как проект TEI стартовал в 1988 г. [1].

Система TEI дает рекомендации по электронной публикации текстов (идентификация текста, представление, анализ и интерпретация, метаязык описания и кодировки). Она в основном рассчитана на текстовые документы, но также предоставляет возможность описания и идентификации других форматов данных, например, графики и звуковых материалов. Главная цель проекта – разработка форматов для обмена данными в гуманитарной области.

---

[1] TEI P4: An XML Version of TEI Guidelines. – http://www.tei-c.org/P4X/AB.html#ABTEI

Рекомендации TEI призваны:

- определить единый синтаксис формата;
- определить метаязык для описания схем представления и кодирования данных;
- описать существующие схемы кодирования на данном метаязыке;
- предложить множество схем описания для разных данных и разных задач;
- обеспечить максимальную совместимость с существующими стандартами;
- поддерживать конверсию схем кодирования существующих машиночитаемых текстов в синтаксис нового формата без добавления какой-либо новой информации в эти тексты;
- обеспечивать возможность использования без специального программного обеспечения.

TEI поддерживают такие международные организации, как Association for Computers and the Humanities (Ассоциация по компьютерам и гуманитарным наукам), Association for Computational Linguistics (Ассоциация по вычислительной лингвистике) и Association for Literary and Linguistic Computing (Ассоциация по компьютерным технологиям в литературе и лингвистике).

Для определения схемы кодирования в TEI используются языки SGML и XMЛ, позволяющие формально определить схему кодирования в терминах элементов и атрибутов, а также с помощью правил, управляющих их размещением в тексте.

Все метки TEI применительно к корпусам можно отнести к различным группам, в частности: метаданные, структурные элементы текста, специальная (лингвистическая) метаинформация.

В TEI языковыми корпусами называются *составные корпусы, т.е. единые цельности, состоящие из множества текстов.* Это объясняется тем, что, хотя каждый отдельный фрагмент текста в корпусе имеет право считаться самостоятельным текстом, в научных целях каждый фрагмент рассматривается также как составляющая большего объекта. Корпусы и другие типы составных текстов (например, антологии и сборники) имеют много общего. Примечательно, что разные компоненты составных текстов могут иметь разные структурные характеристики (например, допускается объединение в корпусе стихов и прозаических текстов), при этом разные компоненты обслуживаются элементами разных модулей TEI.

Помимо основных тегов TEI предлагается ряд специализированных наборов тегов для работы с корпусами.

Рассмотрим основные теги и возможности стандарта с точки зрения многообразия типов корпусов и решаемых в корпусной лингвистике задач.

Для организации основных уровней корпусов предназначены следующие теги:

<teiCorpus> – содержит весь корпус, закодированный в формате TEI; корпус состоит из заголовочного тега корпуса и одного или нескольких тегов TEI, каждый из которых содержит заголовочный тег текста и сам текст.

<TEI> (документ TEI) – содержит один документ, совместимый с форматом TEI; этот документ состоит из заголовочного тега TEI и текста, который располагается изолированно или внутри тега <teiCorpus>.

<teiHeader> (заголовочный тег TEI Header) – содержит описание текста и информацию о его декларации в виде электронной страницы, которая располагается перед началом каждого текста, совместимого с форматом TEI.

@type – указывает на тип документа, к которому относится данный заголовочный тег (является ли документ корпусом или отдельным текстом).

<text> – содержит один текст любого типа, цельный или составной, например, поэму или пьесу, цикл эссе, роман, словарь или фрагмент корпуса.

<group> – содержит сам составной текст, который состоит из различных текстов (групп текстов), которые по какой-то причине рассматриваются как единое целое, например, тексты одного автора, стихотворный цикл и т. д.

Особо следует отметить проработку в TEI разметки корпусов устной речи. Внутри тега **<profileDesc>** может находиться тег **<particDesc>**, который обслуживает дополнительную информацию о говорящих или, если это нужно, о лицах, упомянутых или обсуждаемых в письменном тексте. Нужно отметить, что, хотя употребляются термин *участник речевого акта,* но подразумевается, что все существа, наделенные голосом, в тексте описываются по той же схеме, если не оговорено иное. Идентифицированный персонаж пьесы или романа может считаться полноправным участником речевого акта.

Если в шаблон добавлены элементы модуля **namesdates** (см. тип «*Имена, даты, люди, места»*), внутри тега **<particDesc>** может содержаться подробная информация о говорящем или группе говорящих,

например их имена и другие индивидуальные характеристики. Когда личность говорящего распознана, ему можно присвоить код, которым говорящий будет обозначаться в любом куске кодированного текста, например как определяемый элемент атрибута who. Атрибут who содержит индивидуальные характеристики одного или нескольких участников.

Тег **<settingDesc>** используется для того, чтобы указать, в какой окружающей обстановке происходит речевой акт. Описание окружающей обстановки может быть связным нетегированным текстом (как описание оформления сцены перед началом спектакля). Оно же может быть подробным и тегированным.

Если фигурирует несколько описаний окружающей обстановки, используется несколько тегов <setting>.

<setting> – содержит подробное описание окружающей обстановки, в которой происходит речевой акт.

Если участники речевого взаимодействия находятся в разных местах, то с помощью факультативного атрибута who (реализуемого в теге <setting> как и в любом теге метода att.ascribed), разным участникам могут быть приписаны описания разных окружающих обстановок.

Перечисленные классы для речевой ситуации реализуются с помощью следующих тегов:

<name> (имя собственое) – содержит имя собственное или его транспонированный аналог.

<date> – содержит дату (в любом формате).

<time> – содержит фразу, указывающую на время дня (в любом формате).

<locale> – содержит краткое нетегированное описание места, где происходит речевой акт: в комнате, в ресторане, на скамейке в парке и т. д.

<activity> – содержит краткое нетегированное описание того, чем участник речевого акта занимается во время речевого акта (если он чем-то занимается).

Метаинформация в стандарте TEI получила название контекстуальной информации. Примерами ее служат: возраст, пол и географическое происхождение участников речевого акта, их социально-экономический статус; стоимость и дата публикации газеты; общая тематика или выходные данные книги и т. п. Информация такого рода обладает первостепенной важностью для корпусной лингвистики. Она является организующим принципом при создании корпуса (как в том случае, когда нужно проверить, что с точки зрения некоторой характе-

ристики размах выборки равномерно представлен во всем корпусе или представлен пропорционально численности фрагментов, взятых для составления корпуса), она является критерием выбора фрагментов при поиске и при анализе корпуса (как в том случае, когда требуется изучить специфические языковые характеристики применительно к некоторому сообществу или подмножеству текстов).

Эта информация должна быть зафиксирована в соответствующем разделе заголовочного тега TEI. Метаинформация обо всех документах представлена в отдельном файле с целью удобства выбора подмножества корпуса по определенным признакам.

Тег метаописания документа **<teiHeader>** имеет следующие атрибуты:

- id – уникальный идентификатор документа в корпусе (сейчас он соответствует имени файла без расширения; учитывая, что он составляет основу для идентификаторов слов и предложений, можно его сократить до уникального короткого имени);
- target – имя файла, в котором находится документ;
- type='text' – тип описания, у нас всегда "text", могут быть описания групп документов;
- lang='ru' – язык, на котором написан документ, у нас всегда "ru", в TEI используется указание языка по стандарту ISO 639 (атрибут lang задает значение *по умолчанию*. Это значение может быть переопределено для отдельного предложения или слова, если в русский текст включен фрагмент на другом языке, в TEI предусмотрен также тег <foreign> для иноязыных вставок).

Все метаописание документа состоит из следующих групп элементов:

- **<fileDesc> –** информация о тексте документа;
- **<profileDesc> –** информация о жанре документа;
- **<encodingDesc> –** информация о структуре разметки документа (либо ссылка на стандартную);
- **<revisionDesc>** – информация об истории модификации документа;

Кроме <fileDesc> нам может быть полезен **<profileDesc>**, который содержит информацию об общем классе текстов, например, художественная литература, публицистика, устная речь и т.п.

Описание файла <fileDesc> состоит из следующих элементов:

- <titleStmt>– библиографическая информация о тексте;
- <publicationStmt> – библиографическая информация об издании;

- <sourceDesc> – информация об источнике, из которого получена электронная версия документа.

Библиографическая информация <titleStmt> включает элементы:

- <title> – название;
- <author> – автор;
- <date> – дата создания оригинального документа;
- <extent> – размер документа в некоторых условных единицах (их типология может быть задана в атрибуте type, но для нас естественно считать в словах; надо сформулировать правила для подсчета слов, например, можно считать словом последовательность символов от пробела до пробела, можно, наоборот, только последовательности букв из кирилицы/латиницы, можно только из кириллицы, можно считать многословные, например, *так как, как-нибудь, Нью-Йорк, друг друга* за одно слово; учитывая, что корпус оценивается, в том числе и по длине в словах, требуется точное указание параметра);
- <sponsor> – удобный элемент, в котором мы можем сослаться на соответствующего спонсора (TEI задает еще элемент <funder>, который непонятно чем отличается);
- <respStmt> – информация о человеке/людях, внесших интеллектуальный вклад в создание данного электронного документа (не авторы и спонсоры); <respStmt> задает информацию с помощью элементов <name> и <resp> для указания природы интеллектуального вклада, в нашем случае мы можем вносить сюда ответственных за ручную разметку документа.

2.2.4.3. Лингвистическая разметка

Форматы морфологической разметки. Следует различать формат структуры данных и формат наполнения. С точки зрения структуры, можно выделить три основных способа разметки текста лингвистической информацией:

- простое добавление: за каждым словом следует краткое описание его признаков, например, gives_VVZ, где код VVZ означает, что это третье лицо ед. ч. (Z) значимого глагола (VV) (см. п. 2.3.2, морфологическая разметка Брауновского корпуса);
- набор строк (или таблица): где каждая строка (или строка таблицы) содержит слово и его грамматическую информацию (так называемый вертикальный формат);

- язык разметки: набор средств для записи лингвистической информации, оформленный в виде языка ключевых слов со своим синтаксисом (как правило, языки SGML и XML, см. п. 2.3.2., рис. 1).

В вертикальном формате каждое слово и вся грамматическая информация к нему приводятся отдельной строкой (или строкой таблицы). В этом формате лингвистические параметры часто даются в позиционной системе кодирования, где каждой позиции соответствует определенное грамматическое значение. Примером такой разметки является разметка данных Чешского национального корпуса.

Возможны и комбинированные варианты записи.

Окончательно стандарты наполнения для корпусов еще не сложились. Среди конкурирующих друг с другом стандартов наиболее значимыми являются: EAGLES (European Advisory Group on Language Engineering Standards), TEI (Text Encoding for Interchange), и XCES (XML Corpus Encoding Standard).

Правила EAGLES (EAGLES 1996) задают общие принципы создания и документирования корпусов и их морфосинтаксической разметки, а также ряд конкретных решений для разметки определенных случаев. В частности, они рекомендуют проводить лемматизацию. EAGLES также предлагает две возможности для хранения морфологической разметки: каждый признак представлен отдельным атрибутом (POS='NN' number='sing'), или можно использовать сложную морфологическую аннотацию, в которой цифры соответствуют признакам, например, feats="V3011141101200" означает: глагол, 3rd person, singular, finite, indicative, past tense, active, main verb, non-phrasal, non-reflexive form of a verb (список рекомендуемых признаков и их значений является частью рекомендаций EAGLES). Однако правила EAGLES не содержат готового набора тегов для создания корпуса.

Существующие корпусы, лингвистическая разметка которых основана на SGML/XML, имеют самые разные системы кодирования. Например, в BNC используется CDIF, основанный на TEI; American National Corpus, Croatian National Corpus и др. основываются на XCES; корпусы ICE (International Corpus of English), Czech National Corpus, Hungarian National Corpus и др. ориентируются на стандарт TEI.

Для русского языка стандарт TEI был адаптирован С.А. Шаровым и С.О. Савчук и использован при создании Национального корпуса русского языка.

Наиболее разработанным стандартом для собственно лингвистической разметки текстов является XCES, который также планируется превратить в международный стандарт в рамках проекта ISO TC37/SC4. XCES задает абстрактную *мета*модель, которая обеспечивает средства создания всех разумных моделей лингвистических разметок, удовлетворяющих правилам EAGLES. Для этого определены абстрактные теги узлов <struct> и их признаков <feat>. Для каждого узла должен быть задан его тип, например, p-level, s-level, w-level, m-level, соответственно, для абзацев, предложений, слов и морфем. Это позволяет представлять мультислова как одну единицу анализа, например, as well as в английском или глаголы с отделяемыми приставками, например zunehmen в немецком. Можно также проводить декомпозицию одного слова в пределах разметки, например, для zum как zu dem в немецком.

В качестве одного из стандартов морфологической разметки следует назвать многоязыковые морфосинтаксические спецификации (multilingual morphosyntactic specifications) MULTEXT-East Version 4 (http://nl.ijs.si/ME/V4/).

Форматы кодирования синтаксических отношений. Достаточно широк набор языков для синтаксической разметки текстов. Например, язык горизонтальной записи корпуса PennTreebank основан на хранении деревьев в виде LISP-списков:

```
(S      (NP-SB (PPH-HD He))
        (VP-OC (VVD-HD studied)
                (NP-DO (ART-ND the) (NN-HD problem))...
```

В TEI для синтаксических отношений имеются стандартные теги:

- тег <cl> – клауза, для кодирования сложносочиненных и подчиненных предложений, у него есть два атрибута – type, задающий ее синтаксические признаки, и function, задающий ее функцию;
- тег <phr> – группа, аналогично атрибут type задает ее тип (именная, предложная и др.), и function задает ее функцию.

Для представления в терминах зависимостей можно предусмотреть специальные теги, например, <depc>, который имеет атрибуты function и target, последний ссылается на идентификатор зависимого слова в предложении.

Приведем пример морфосинтаксической разметки в TEI:

```
<p>
<cl type='finite declarative' function='independent'>
```

```
<phr type='NP' function='subject'>Nineteen fifty-four,
<cl type='finite relative declarative'
 function='appositive'>when
 <phr type='NP' function='subject'>I</phr>
 <phr type='VP' function='predicate'>was eighteen
      years old</phr>
</cl>,
</phr>...
```

Вопросы для самоконтроля

1) Дайте определение следующих понятий: *корпусный менеджер*, *лемматизация*, *метаданные*, *парсинг*, *стемминг*, *токен*.

2) Дать понятие разметки.

3) Какие типы лингвистической разметки существуют?

4) Что представляет собой экстралингвистическая разметка?

5) Каковы подходы к стандартизации форматов представления данных с точки зрения их наполнения и структуры?

6) Перечислить и характеризовать основные процедуры обработки естественного языка при создании корпусов.

Часть 3
ИСПОЛЬЗОВАНИЕ КОРПУСОВ

3.1. Корпусные менеджеры

3.1.1. Корпус как поисковая система

Неотъемлемой частью понятия «корпус текстов» является ***корпусный менеджер*** – специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме.

Корпусный менеджер должен:

- строить как KWIC, так и полные конкордансные списки;
- искать контексты не только по отдельным словам, но и по словосочетаниям;
- осуществлять поиск по шаблонам (сложные запросы);
- сортировать полученные списки по нескольким критериям, выбираемым пользователем;
- давать возможность отображать найденные словоформы в широком контексте;
- давать статистическую информацию по отдельным элементам корпуса;
- отображать леммы, морфологические характеристики словоформ и метаданные (библиографические, типологические), что зависит от степени размеченности корпуса;
- сохранять и распечатывать результаты;
- работать как с корпусами (неограниченными по размеру), так и с подкорпусами;
- поддерживать различные форматы текстовых данных (txt, doc, rtf, html, xml и др.);
- быть легким (интуитивно понятным) в использовании, как для опытного, так и для начинающего пользователя.

Наиболее известны такие универсальные корпусные менеджеры, как SARA, XAIRA (BNC), Manatee/Bonito, CQP, DDC. Для обработки корпусных данных могут разрабатываться менеджеры на основе систем управления базами данных (СУБД) или поисковых систем. Например, поиск по Национальному корпусу русского языка осуществляется поисковой системой Яndex.Server 3.8 Professional [27].

3.1.2. Языки запросов

Информационный запрос – это словесное выражение определенной информационной потребности. Запросы анализируются по своему предметному и формальному содержанию и описываются в терминах языка запросов прикладной программы, работающей с корпусом. Процедура поиска заключается в сопоставлении поискового образа запроса с отдельными элементами данных корпуса и в вычислении их соответствия.

Далее будет рассмотрен язык запросов одного из наиболее эффективных корпусных менеджеров, **Bonito/Manatee**[1]. На примере этой поисковой системы будет продемонстрировано большинство основных элементов языка запросов к корпусам текстов, а также приведены примеры задания запросов к корпусу.

Корпусный менеджер Bonito представляет собой программное обеспечение для работы с корпусами текстов. Система Bonito состоит из двух частей: сервера (Bonitosrv) и графического пользовательского интерфейса (GUI – graphical user interface) Bonito, созданных П. Рыхли и группой NLPlab (Natural Language Processing Laboratory) на факультете информатики Университета им. Масарика (Чехия).

Для демонстрации работы с системой будет использоваться корпус английских текстов SUSANNE (Surface and Underlying Structural Analysis of Natural English) (http://www.grsampson.net/). Данный корпус был создан в Великобритании в Университете Сассекса. Он включает в себя более 130 тыс. слов Брауновского корпуса, аннотированного согласно схеме SUSANNE.

Основные особенности системы Bonito

Язык запросов

- поиск отдельных атрибутов (словоформа, лемма, тэг);

[1] Bonito – название менеджера, Manatee – вся программная подсистема корпусного обеспечения. Современная версия системы называется Nosketch Engine.

- использование регулярных выражений;
- логические операторы;
- средства задания структуры (границы предложения и др.);
- быстрая обработка сложных запросов;
- шаблоны.

Конкордансные списки

- история запросов пользователя;
- просмотр морфологических характеристик словоформы;
- отображение леммы.

Операции над конкордансом

- сохранение списков в файл;
- печать списков;
- сортировка по ключевым словам, контексту;
- интерактивное неограниченное расширение контекста;
- фильтрация (удаление части построенных конкордансов);
- удаление повторений.

Частотное распределение

- частоты слов и других атрибутов в корпусе, контексте;
- неограниченное число уровней группировки.

Другие особенности

- выбор кодировок;
- создание пользовательских подкорпусов;
- произвольный набор тэгов;
- возможность подключения других языков.

Запросы

Пользователь может ввести собственно запрос, сформулированный по правилам языка запросов системы, или шаблон (готовый или созданный пользователем) в окно запросов (рис. 3).

Типы запросов:

Положительный фильтр (P-filter) – совпадающие с запросом строки выдаются в конкордансном списке;

Отрицательный фильтр (N-filter) – совпадающие с запросом строки удаляются из конкордансного списка;

Словосочетания (Collocations) – удовлетворяющие запросу позиции (конкретная словоформа на заданном интервале) в конкордансе выделяются цветом.

Для положительного, отрицательного фильтров и словосочетаний необходимо задавать интервал, в пределах которого следует искать совпадающие позиции для каждой строки конкорданса. Пользователь

задает границы интервала (окна ввода "**From:**" и "**To:**"). Если значения положительные, то поиск организуется вправо от исходной позиции, если отрицательные – то влево.

Исходной позицией может служить начало словоформы, конец словоформы, начало N-й позиции, конец N-й позиции. Очень важно отметить, что все введенные запросы сохраняются в так называемой Истории запросов (Query History), но если запрос идентичен одному из предыдущих, он не попадает в Историю запросов.

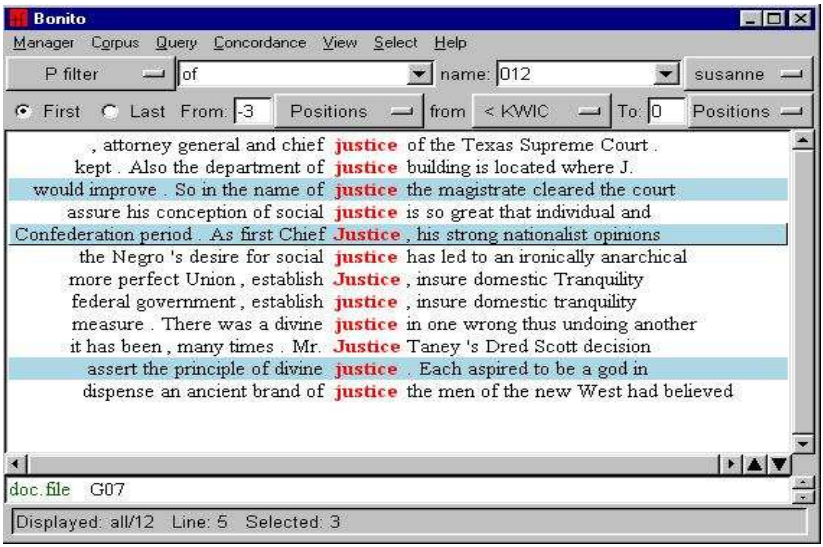


Рис. 3. Окно корпус-менеджера Bonito с конкордансом для словоформы "justice".

Достаточно нажать стрелку «вниз» в окне запроса, чтобы проследить всю Историю, а если необходимо, то вернуться к одному из предыдущих введенных запросов. Если ввести имя запроса в окне "name", то запрос сохраняется в списке поименованных запросов (named queries).

Шаблоны

Шаблон – это вид запроса, который упрощает ввод однотипных запросов. Это означает, что сложный запрос необходимо создать только один раз и сохранить как шаблон, а затем просто вводить значения для данного шаблона.

Например, шаблон для всех словоформ правильного английского глагола play мог бы выглядеть так:

[word="$1" | word="$1s" | word="$1ed" | word="$1ing"]

В этом шаблоне использовалась переменная, состоящая из значка "$" и цифры "1". Количество переменных в шаблоне не ограничено. При использовании шаблона первый вводимый параметр соответствует переменной $1, второй – $2 и т. д. Параметры вводятся через пробел.

Когда шаблон активизируется, он автоматически записывается в окно запроса. Отличие от обычного запроса состоит лишь в следующем: первый знак строки – это восклицательный знак (!), далее идет имя шаблона, двоеточие (:) и параметры, разделяемые пробелами. Если бы имя приведенного выше шаблона было "regular verb", то строка запроса для всех форм глагола play выглядела бы так:

!regular verb: play

Примеры запросов

В приведенных ниже примерах наглядно продемонстрированы элементы языка запросов корпусного менеджера Bonito.

Пример 1. Поиск конкретной словоформы.

В окно запроса вводится словоформа "run". Выдается:

announced that he would not **<run>** for reelection . Georgia
medical benefits paid out would **<run>** 1 billion or more in the
May , said today Jones will **<run>** well ahead of his GOP opponents
reports that he had decided to **<run>** and wanted Mr. Screvane ,
investigation Street car tracks **<run>** down the center of Pennsylvania

Система ищет полное соответствие запрашиваемому слову и выдает результат. Иных словоформ для конкретной словоформы "run" не будет найдено.

Пример 2. Поиск синтагмы.

Допустим, нужно найти разрывную синтагму "take (smth) out".

В окно запроса вводится "take". Строится конкорданс для данной конкретной словоформы. Выбирается тип запроса: Положительный фильтр (P-filter). В оба окна "From:" и "To:" вводится значение "2", что соответствует второй позиции справа от найденного слова для "оторванной" части синтагмы (в нашем примере "out"). В окно запроса вводим "out". Выдается:

for governor would force it to **<take>** petitions **out** into voting
the peasant . Nonetheless , they **<take>** time **out** -- much time --
Mister McBride . You do that or **<take>** you **out** a permit right now

Разумеется, можно придумать и более сложные варианты подобных запросов с неоднократным применением Положительного фильтра.

Пример 3. Поиск различных форм слова.

В окно запроса вводится "runs? in".

В данном запросе используется *управляющий символ* "?", который означает, что предшествующая ему буква "s" может встретиться ноль или один раз. Полученный результат подтверждает это. Выдается:

tied the game , and single **<runs in>** the eighth and ninth gave
record in the 600 - yard **<run in>** the Knights of Columbus track
their eight hits for two **<runs in>** the sixth . Chuck Hinton
The Bears added their last **<run in>** the sixth on Alusik 's double
's first major league home **<run in>** the fifth put the Sox back

Пример 4. Поиск различных форм слова.

В окно запроса вводится "run(|s|ning)".

Здесь используются группирующие скобки и оператор альтернативы (|) (логическое «или»). Системе дается команда найти конкретные словоформы "run" или "runs" или "running". Выдается:

announced that he would not **<run>** for reelection . Georgia
medical benefits paid out would **<run>** 1 billion or more in the
the group are interested in **<running>** on the required non -
lawyer and former FBI man is **<running>** against the Republican

Пример 5. Поиск всех форм слова по лемме.

В окно запроса вводится "[lemma="be"] within <head>". Выдается:

<head>DECISIONS **<ARE>** MADE</head>Asked to elaborate
<head>LEADERSHIP **<IS>** HOPEFUL</head>The housing
Nations .<head>FORMULA **<IS>** DUE THIS WEEK</head>The
year .<head>COULD **<BE>** SCRAMBLE</head>Some predict
ends .<head>CHOICE **<WAS>** EXPECTED</head>The selection
<head>TOBACCO ROAD **<IS>** DEAD . LONG LIVE TOBACCO

Так можно не только искать все словоформы по лемме, но и находить их в заданных полях документа (в данном примере в заголовочном поле, обозначенном тэгом <head>). Соответственно, если ввести несколько лемм подряд, то можно получить все варианты таких словосочетаний.

Пример 6. Поиск по морфологическим признакам.

В окно запроса вводится "[tag="VVZv"]". Выдается:

charge of the election , " **<deserves>** the praise and thanks of the

However, the jury said it **<believes>** " these two offices should be
of Fulton County , which **<receives>** none of this money " . The
when the new management **<takes>** charge Jan. 1 the airport be
face is a state law which **<says>** that before making a first

Пример демонстрирует замечательную возможность корпусного менеджера искать словоформы по морфологическим признакам. Код "VVZv" означает, что это – третье лицо единственного числа (Zv) значимого глагола (VV). Такая кодировка предложена схемой аннотирования SUSANNE. Следовательно, данная возможность будет успешно использоваться, в первую очередь, теми, кто знаком с принципами данной схемы аннотирования.

Пример 7. Отображение морфологических признаков и леммы.

В пункте командного меню выбирается "**View ⇒ Attributes…**" и отмечаются пункты "**lemma**" и "**tag**".

В окно запроса вводится "[lemma="be"]". Выдается:

in which the election **<was/**be/VBDZ> conducted . The September
October term jury had **<been**/be/VBN> charged by Fulton Superior
and election laws " **<are/**be/VBR> outmoded or inadequate
these two offices should **<be/**be/VB0> combined to achieve greater
Department, the jury said, " **<is/**be/VBZ> lacking in experienced

В конкордансе для каждого вхождения словоформы показана ее исходная форма и ряд морфологических признаков в виде кода.

Язык регулярных выражений RegEx

Языки запросов корпусных менеджеров, представленные в той или иной форме (формализованный язык запросов или оконный интерфейс), как правило, базируются на формализме, который получил название «язык регулярных выражений». Бóльшую часть запросов на языке RegEx «скрывают» от пользователя в программном коде, реализовав их в виде удобного интерфейса. Пользователю необходимо лишь заполнить определенные поля формы (web-страница с ячейками для заполнения), и его запрос будет осуществлен. Но все же для сложных запросов полезно знать основы языка регулярных выражений.

Регулярные выражения – это строковые записи, задающие правила поиска на особом языке. Если есть выражение и какая-либо строка (слово, массив текстов, записи в полях базы данных и т. д.), то операцию проверки, удовлетворяет ли строка выражению, называют *сопоставлением* строки и выражения. Если какая-то строка или часть строки успешно сопоставилась с выражением, это называется *совпадением*

(соответствием). Например, при сопоставлении выражения «группа букв, окруженная пробелами» и строки «*помню чудное мгновенье*» совпадением будет строка «*чудное*» (ведь только она удовлетворяет данному выражению).

Существует несколько разновидностей языков, используемых для записи регулярных выражений и работы с ними. Так, в популярном языке программирования PHP4 и СУБД MySQL реализован язык регулярных выражений RegEx. У них есть много общего, но отдельные части все же отличаются.

В языке RegEx каждое выражение состоит из одной или нескольких управляющих команд. Некоторые из них можно группировать, и тогда они принимаются за одну команду. Все управляющие команды разбиваются на три класса:

1) *простые символы*, а также *управляющие символы*, играющие роль их заменителей;

2) *управляющие конструкции* (квантификаторы повторений, оператор альтернативы, группирующие скобки и т. д.);

3) так называемые *мнимые символы* (в строке их нет, но они «помечают» какую-то часть строки, например, ее конец).

Простые символы. Класс простых символов, действительно, самый простой. А именно, любой символ в строке на языке RegEx обозначает сам себя, если он не является управляющим. К управляющим символам причисляются следующие: .*?+[]{}|$^

Например, регулярное выражение "abcd" будет «реагировать» на строки, в которых встретится последовательность "abcd".

Группы символов. Одним из самых важных управляющих символов является точка "**.**", обозначающая один любой символ. Например, выражение "л.к" имеет совпадение для строк "лик", "лук", "лак". Позже будет показано, как можно с помощью точки обозначить ровно один любой символ (или, например, ровно пять).

Возможно, понадобится искать не любой символ, а один из нескольких указанных. Для этого нужно заключить их в квадратные скобки. К примеру, выражение "л[иуа]к" соответствует строкам, в которых есть подстроки из трех символов, начинающиеся с "л", затем одной из букв "и,у,а" и, наконец, "к". Если букв-альтернатив много и они идут подряд (в алфавитном порядке), то не обязательно перечислять их все. Достаточно указать через дефис первую и последнюю. Например, выражение "[а-я]" обозначает любую букву от "а" до "я", а выражение "[а-я0-9]" задает любой алфавитно-цифровой символ.

Существует и другой, иногда более удобный способ задания больших групп символов. В языке RegEx в квадратных скобках могут встречаться специальные выражения, обозначающие сразу группу символов:

- [:alpha:] – буква;
- [:digit:] – цифра;
- [:alnum:] – буква или цифра;
- [:space:] – пробельный символ;
- [:punct:] – знак пунктуации.

Отрицательные группы. Иногда, когда альтернативных символов много, бывает довольно утомительно перечислять их все в квадратных скобках, особенно если подходят все символы, кроме нескольких. В этом случае следует воспользоваться конструкцией "[^]", которая обозначает любой символ, кроме тех, что перечислены после "[^" и до "]". Например, выражение "м[^ао]х" будет соответствовать всем строкам, содержащим буквы "м" и "х", разделенные любым символом, кроме "а" или "о".

Управляющие конструкции. Квантификаторы повторений. Перейдем к рассмотрению так называемых квантификаторов – спецсимволов, использующихся для уточнения действия предшествующих им символов первого класса.

Ноль и более совпадений. Звездочка "*" обозначает, что предыдущий символ может быть повторен ноль или более раз. Например, выражение "19*8" соответствует строке, в которой есть цифра "1", затем ничего или несколько цифр "9" и, наконец, цифра "8".

Одно и более совпадений. Символ плюса "+" обозначает одно или более совпадений предшествующего символа или группы. Вот пример выражения, которое определяет слова, написанные через дефис**:** "[а-я]+-[а-я]+".

Ноль или одно совпадение. Иногда используют еще один квантификатор – знак вопроса "?". Он обозначает, что предыдущий символ может быть повторен ноль или один (но не более!) раз. Например, выражению "Петров[аы]?" будут соответствовать строки "Петров", "Петрова" и "Петровы".

Заданное число совпадений. Последний квантификатор повторения – фигурные скобки "{}". С его помощью можно реализовать все перечисленные выше возможности. Существует несколько форматов его записи:

- A{n,m} – указывает, что символ "A" может быть повторен от n до m раз;
- A{n} – символ "A" должен быть повторен ровно n раз;
- A{n, } – символ "A" может быть повторен n или более раз.

Оператор альтернативы. При описании простых символов была рассмотрена конструкция "[…]", которая позволяла указывать, что в нужном месте строки должен стоять один из указанных символов. Это не что иное, как оператор альтернативы, работающий с отдельными символами.

В языке RegEx есть возможность задавать альтернативы не одиночных символов, а сразу их групп. Это делается при помощи оператора "|". Вот несколько примеров его работы:

- "1|2|3" – полностью эквивалентно выражению [123];
- "^пре|^пере" – строки, которые начинаются с "пре" или "пере";
- "давать|давал|давала|давало|давали" – соответствует подстрокам, разделенным символом альтернативы "|".

Группирующие скобки. В примере "давать|давал|давала| давало| давали" подстрока "дава" встретилась в выражении пять раз. Для управления оператором альтернативы существуют группирующие круглые скобки "()". С их помощью выражение из последнего примера можно было записать так: "дава(ть|л|ла|ло|ли)". Скобки могут иметь произвольный уровень вложенности.

Мнимые символы. Мнимые символы – это просто участок строки между соседними символами, удовлетворяющий некоторым свойствам. Фактически, мнимый символ – это некая позиция в строке. Так, символ "^" соответствует началу строки, а "$" – ее концу.

Например, выражение "^пере" будет соответствовать любой строке, начинающейся на "пере", выражение "ть$" – строке, оканчивающейся на "ть", а выражение "^перенять$" – точному совпадению со строкой "перенять".

3.1.3. Выходные интерфейсы

Результаты поиска (выдача) в корпусных менеджерах обычно представлены в виде конкорданса.

Согласно словарю иностранных слов ***конкорданс*** – это расположенный в алфавитном порядке перечень встречающихся в книге слов с минимальным контекстом (в несколько слов). В словаре Collins Cobuild English Dictionary слово **concordance** определяется следующим обра-

зом: an alphabetical list of the words in a book or a set of books which also says where each word can be found and often how it is used.

В корпусной лингвистике – это список всех употреблений заданного в результате поиска языкового выражения (обычно слова) в контексте, возможно, со ссылками на источник.

Ниже приведен фрагмент конкорданса для слова «имение» из текста «Дубровский» А.С. Пушкина (рис. 4).

с в губерниях, где находилось его	**имение**	. Соседи рады были угождать мал
грубиян; я хочу взять у него	**имение**	, как ты про то думаешь? – Ваше
чтобы безо всякого права отнять	**имение**	, Постой однако ж. Это имение
имение. Постой однако ж. Это	**имение**	. принадлежало некогда нам, было

Рис. 4. Конкорданс KWIC для слова «имение».

Распространенный подход к показу контекстного окружения состоит также в переходе от формы конкорданса (рис. 4) к широкому контексту (рис. 5).

View 1 examples

id=http://piligrim.iatp.by/article.html **title="Мирский замок"** уже не существует &; quot;. Последовало еще несколько писем в газету по поводу разрушения замка, и лишь после этого разборка его на кирпич прекратилась. Только через десять лет были, наконец, накрыты четыре башни Мирского замка гонтовыми крышами. Последующие владельцы – радзивилловские отпрыски, породненные с немецкой фамилией Гогенлое-Шиллингфюрст, совершенно не интересовались ни поселком, ни замком, сдавая **имение** в аренду. Один из арендаторов, некий Антон Путята, занялся устройством вокруг замка фруктовых и цветочных садов, но дело закончилось полным разорением предпринимателя и окончательным запустением замка. К этому времени относятся первые печатные изображения Мирского замка. В книге Ф. М. Собещанского, изданной в Варшаве на польском языке в 1849 году, вместе с рисунком замка появилось и первое его описание как памятника

Рис. 5. Расширенная форма выдачи корпусного менеджера CQP.

И конкорданс KWIC (Key Words In Context), и более полный контекст могут оказаться полезными в зависимости от того, для каких целей нужен данный материал.

Существуют специальные программы составления конкордансов по некоторому корпусу текстов – конкордансеры. Они позволяют получать список контекстов, в которых заданная единица встретилась, ее частоту, позволяют сортировать контексты по ключевому слову, по ближайшему контексту. Более сложные программы (корпусные менеджеры) способны строить полные конкордансы, включающие в себя не только слова, но и другие элементы корпуса.

Имеется множество параметров, которые бывает необходимо получить из корпуса. Это лемма и морфологические характеристики слова; позиция слова в предложении и в структуре размеченного текста (HTML, XML); библиографические и типологические признаки документа, из которого выбран контекст (автор, название, источник, год издания, тип текста и т. д.); статистические данные и многое другое.

Обработка конкордансных данных с учетом статистических сведений о лексических единицах корпуса позволяет вычислять силу синтагматической связанности между лексемами. Подробнее см. п. 3.3.3.2.

3.1.4. Корпусные менеджеры нелингвистических корпусов (поисковые системы Интернета)

Информационное наполнение сети Интернет (веб-пространство) может рассматриваться как огромный многоязычный корпус. Главный материал лингвистического анализа – язык, зафиксированный в виде речевых произведений, – в Интернете представлен в огромном объеме и разнообразии и непосредственно доступен для машинной обработки. Этот факт представляет для лингвистов большую ценность, так как перевод текстов в машинную форму и создание корпусов требует больших временны*х* и материальных затрат.

Текстовые массивы Интернета широко используются как источник данных для формирования корпусов. Например, в 2011 г. на сайте http://corpus.byu.edu был размещен Google Books (American English) Corpus, объемом 155 млрд слов, основанный на данных Google Books и включающий тексты книг на американском варианте английского языка с 1810 по 2009 гг. Так же широко тексты, представленные в Интернете, используются как тестовый материал для различных программ анализа и обработки текстовой информации (особенно тех, которые базируются на статистических и вероятностных методах).

В то же время веб-пространство может рассматриваться и непосредственно как корпус. Особенно активно эта проблема стала обсуждаться после доклада А. Килгарриффа в 2001 г. [45]. Очевидно, что ни один корпус не может сравниться по репрезентативности языкового материала с вебом, куда включаются материалы и других Интернет-сервисов (например, электронной почты). При этом, однако, встает вопрос о сбалансированности такого веб-корпуса. Очевидно, что в Интернете определенные типы речевых произведений представлены относительно чаще, чем это было в языке до сих пор.

При использовании веб-пространства как корпуса роль корпусных менеджеров могут выполнять поисковые системы. Основным средством поиска информации в сети являются глобальные информационные поисковые системы вербального типа (поисковые машины – search engines), индексирующие все Интернет-пространство и обеспечивающие поиск по тексту. При этом полезно представлять, как эти индексы вербальных систем строятся, и, соответственно, учитывать эти особенности при использовании баз данных поисковых систем как материала для лингвистических исследований.

Существует большое количество таких систем, отличающихся друг от друга языком запросов, дизайном, сервисом и другими особенностями.

В составе любой поисковой системы можно выделить три основные части:

1. ***Робот*** – подсистема, обеспечивающая просмотр (сканирование) Интернета и поддержание инвертированного файла (индексной базы данных) в актуальном состоянии. Этот программный комплекс является основным средством сбора информации о наличии и состоянии информационных ресурсов сети.

2. ***Поисковая база данных*** – так называемый ***индекс*** – специальным образом организованная структура данных (*англ.* index database), включающая, прежде всего, инвертированный файл, состоящий из лексических единиц, взятых из проиндексированных веб-документов, и содержащий разнообразную информацию об этих единицах (в частности, их позиции в документах), а также о самих документах и сайтах в целом.

3. ***Поисковая система*** – подсистема поиска, обеспечивающая обработку запроса (поискового предписания) пользователя, поиск в базе данных и выдачу результатов поиска пользователю. Поисковая система общается с пользователем через пользовательские интерфейсы – экранные

формы программ-браузеров: интерфейс формирования запросов и интерфейс просмотра результатов поиска.

Фактически индексы (инвертированные файлы) поисковых систем – это, по сути, не что иное, как виртуальные конкордансы к текстам. Более того, результаты поиска в информационных поисковых системах в виде кратких описаний документов, как правило, содержат контексты, в которых искомые слова встретились в найденных документах. Отличие лишь в том, что конкордансы обычно составляются к конкретному произведению или группе произведений (например, все тексты одного и того же автора), в то время как информационная поисковая система Интернета индексирует все доступное множество электронных документов.

Главная содержательная проблема при индексировании веб-сайтов заключается в том, какие термины попадают в индекс. Активно применяются списки запрещенных слов (stop-words), которые в индекс не попадают – это служебная лексика (предлоги, союзы и т. д.) и незначащие слова.

Многие системы индексируют лишь часть документа (обычно начальную), есть роботы, которые обрабатывают только часть вебстраниц с одного и того же сайта. Знание того, как работают роботы, каковы их технические характеристики, полезно и для создателей вебдокументов, и для составителей запросов при поиске. Подробное описание работы роботов можно найти в Сети[1].

Важно, какую информацию и в каком виде можно извлечь из выходных интерфейсов информационной поисковой системы (ИПС). Интерфейс выдачи (форма представления результатов) у разных систем включает такие параметры, как статистика слов из запроса, количество найденных документов, количество найденных сайтов, количество документов на странице с результатами поиска, средства управления сортировкой документов в выдаче, описание сайта, с которого взят соответствующий документ, описание документа. Последнее, в свою очередь, может содержать в своем составе заглавие документа, URL (адрес в сети), аннотацию (фрагмент текста с выделенными словами из запроса), указание на другие релевантные веб-страницы того же сайта, ссылка на рубрику каталога, к которой относится найденный документ или сайт, коэффициент релевантности, ссылки на другие возможности поиска (поиск похожих документов, поиск в найденном).

---

[1] См., в частности, http://citforum.ru/internet/search/art_1.shtml; http://www.webmasterpro.com.ua/news36.html

Из всех этих реквизитов для задач лингвистического исследования наибольший интерес представляют частотные характеристики и выдача контекста. Следует различать два типа частот, учитываемых и выдаваемых системами – пословную и подокументную. Сведения о количестве языковых единиц в разных системах и разных режимах поиска могут относиться как к словоформам, так и к лексемам. Некоторые системы ведут журнал запросов с возможностью повторных поисков и с выдачей статистики по запросам. Полезной и интересной возможностью является также отнесение документов к тематическим классам [15].

3.2. Обзор существующих корпусов различных типов

3.2.1. Зарубежные национальные корпусы

В настоящее время существуют национальные общеязыковые корпусы для большинства основных языков мира. Остановимся на некоторых из них.

Британский национальный корпус (British National Corpus, BNC) является одним из больших эталонных корпусов, в нем содержится 100 млн слов. Корпус был разработан в Оксфордском университете при участии Ланкастерского университета и Британской библиотеки. Работа над созданием корпуса продолжалась с 1991 по 1994 г. Подкорпус, представляющий письменный английский язык, составляет 90 % всего корпуса и включает в себя художественную и документальную прозу, газеты, периодические научные издания и журналы, издаваемые для различных возрастов, популярную научную фантастику, опубликованные и неопубликованные письма, школьные и университетские сочинения и др.

Корпус включает много разных стилей и не ограничен в тематике. Подкорпус устной речи включает в себя речь добровольно вызвавшихся участвовать в проекте людей различных возрастов, проживающих в разных частях Великобритании и принадлежащих к различным социальным классам. Разговорная речь присутствует в множестве контекстов: от разговоров при формальных деловых или правительственных встречах до радиошоу и телефонных разговоров.

Все тексты Британского национального корпуса сегментированы на предложения. Словам внутри предложения присвоены соответствующие маркеры, обозначающие грамматический класс слова или его часть речи. Знакам препинания тоже были присвоены соответствующие маркеры. Сегментацию и автоматическое присвоение словам тэгов

выполняет программа CLAWS, разработанная в университете Ланкастера. Процент ошибочной разметки составляет примерно 1,7 %. Кроме того, если программа автоматической разметки сталкивалась со случаями, когда она не могла однозначно присвоить слову какой-то маркер, ему присваивались сразу два маркера (например, VVD-VVN – первый обозначает глагол прошедшего времени, а второй – причастие прошедшего времени). Такие «синонимичные» маркеры составляют примерно 4,7 % всего корпуса.

Корпус состоит только из текстов современного английского языка, используемого в Великобритании, однако слова не британского происхождения и иностранные слова, используемые в британском английском, также встречаются в корпусе.

Тексты, представленные в Британском национальном корпусе, отбирались по трем основным критериям: время, область, которую данный текст описывает, и тип издания. По времени все тексты принадлежат примерно одному периоду, начиная с 1975 г., исключения делались только для художественной литературы, поскольку некоторые произведения популярны и по сей день. К области художественной литературы принадлежит 25 % текстов. Литературные произведения в BNC представлены, начиная с 1964 г. 75 % письменных текстов были взяты из информативных изданий (наука, искусство, коммерция и финансы, досуг, социология, политика). Для обеспечения сбалансированности учитывались также размер (количество слов), тема, обсуждаемая в тексте, имя автора, возраст, пол, место рождения, место жительства, возрастная группа людей, которым предназначен данный текст, а также «уровень» сложности данного текста.

Весь 10-миллионный подкорпус устной речи разделен на две примерно равные части: 1) *демографическую* часть, содержащую транскрипции «спонтанных», естественных диалогов, и 2) часть, в которой важную роль играл контекст, так называемую *контекстно-управляемую* часть, содержащую записи, сделанные на каких-либо публичных мероприятиях.

1) *Демографическая* часть. Всего в записи диалогов участвовало 124 добровольца, живущих по всей территории Великобритании, которые должны были носить с собой магнитофоны в течение нескольких дней при выполнении различных действий, фиксируя в записных книжках, при каких условиях состоялись разговоры, и другие моменты – кто являлся собеседниками, каковы были их взаимоотношения, физическое окружение в момент записи речи и т. д. Добровольцы были отобраны так, чтобы было примерно равное количество мужчин и женщин из каж-

дой возрастной группы и из различных социальных классов. У тех, кто принимал участие в записи на пленку, после беседы спрашивали разрешение на то, чтобы их речь была включена в корпус. Затем эти магнитные записи были обработаны, и тексты были записаны обычной английской орфографией. Эти разговоры сейчас используются как основа изучения характера устной речи, и результаты оказываются полезными и интересными [42].

2) *Контекстно-управляемая* часть. Создатели преследовали цель собрать равное количество записей из следующих четырех довольно широких категорий социального контекста:

- образовательные и информативные собрания, такие как лекции, программы новостей, обсуждение чего-либо в классе, семинары;
- деловые события, такие как выставки, консультации, интервью, собрания торговых организаций;
- публичные события, такие как проповедь, политические речи, заседания парламента;
- темы, касающиеся досуга, такие как спортивные комментарии, клубные встречи.

Разработчики создали на основе разметки SGML собственную программу, которую назвали SARA (SGML Aware Retrieval Application). SARA была изначально разработана как программа клиент/ сервер, т. е. система, где один или более компьютеров имеют по сети доступ к центральному серверу. В настоящее время создан новый корпусный менеджер – XAIRA (XML Aware Indexing and Retrieval Architecture).

Одним из наиболее известных корпусов общего типа является ***Чешский национальный корпус*** (Český národní korpus) (далее ЧНК). Это синхронический морфологически размеченный корпус, представляющий современный чешский язык. Созданием корпуса занимается Институт ЧНК под руководством проф. Ф. Чермака. Институт был создан на базе философского факультета Карлова университета в Праге в 1994 г. и функционирует на средства грантов, спонсоров и при поддержке Министерства образования. С работой института и с самим корпусом можно ознакомиться на сайте http://www.korpus.cz.

Первоначальный корпус, насчитывавший 100 млн словоупотреблений письменных текстов, содержал также небольшие коллекции разговорной (750 тыс. словоупотреблений) и диалектной речи. Впоследствии этот корпус, в основной массе состоящий из текстов 1990–1999 гг., получил название SYN2000. Затем были созданы 100-миллионные сбалансированные корпусы SYN2005 (2000–2004 гг.) и

SYN2010 (2005–2009 гг.), а также другие корпусы, прежде всего публицистические. Все синхронические корпусы объединены в общий «пул» объемом 1300 млн словоупотреблений. Кроме того, в составе ЧНК создано несколько корпусов устной (разговорной) речи общим объемом 4 млн словоупотреблений, диахронический корпус (1850 тыс. словоупотреблений, 2005 г.), параллельный корпус (25 языков, 92 млн словоупотреблений, 2008 г.) и др. (http://www.korpus.cz/ struktura.php).

При формировании ЧНК большое внимание уделялось вопросам репрезентативности корпуса. Было принято решение, что основную часть корпуса составят тексты 1990–1999 гг. с дополнительной ретроспективной частью, представляющей собой произведения чешской литературы до 1950 г.

В результате книговедческих исследований была определена жанровая и тематическая структура корпуса, которая выглядит следующим образом (табл. 3).

Таблица 3. Фрагмент жанровой и тематической структуры ЧНК (SYN2000)

Художественные тексты	15 %
Информативные тексты	85 %
в том числе:	
публицистические	60 %
научные	25 %
в том числе:	
социальные науки	3,6 %
естественные науки	3,4 %
искусствоведение	3,4 %
технические науки и т. д.	4,6 %

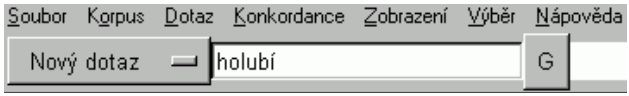
Все тексты хранятся в трех видах: текстовый архив (в исходном виде), банк данных (обработанные тексты на языке SGML) и собственно корпус (тексты в специальном формате и средства работы с ними). Исходные тексты проходят несколько этапов конвертирования, в ходе которых осуществляется их перекодировка (если требуется), структурирование текста, удаление нетекстовых и иноязычных элементов, удаление дублей и собственно разметка. В конечном счете формируется следующая структура: файл, документ, абзац, предложение, словоформа. Каждая структурная единица оформляется как элемент языка SGML. Заголовочная часть каждого файла описывает характеристики процесса конвертирования, заголовочная часть документа содержит библиографические и типологические признаки (автор, название, ис-

точник, год издания, тип текста, жанр и т. п.). Лингвистическая разметка заключается в лемматизации и приписывании словоформам морфологических характеристик, записываемых в позиционном формате как отдельный элемент языка SGML.

Работа с корпусом осуществляется через корпусный менеджер BONITO (http://www.korpus.cz/bonito) (см. п. 3.1.2).

Приведем несколько примеров выдачи из корпуса:

1. Поиск словоупотреблений слова holubí (голубиный).



jej . Hedvábným slunečníkem **holubí** barvy procházelo slunce a
podobné zábrany rovněž slabé a **holubí** samec dokáže v těsné kleci
řeklo by se , že vybírala **holubí** vejce , to se o ní vědělo
odpověděl , " červené jako **holubí** nožky , červenější než ohromné
uniformu tvořila tunika v barvě **holubí** šedi , která se oblékala
znám jitra nadšená jak hejna **holubí** , a časem viděl jsem , co

2. Поиск словосочетания «holubí vejce» (голубиное яйцо) в любой форме и в любом написании (строчные и прописные)

[lemma="holubí"] [lemma="vejce"]

uzavřený v plachetce a připomíná **holubí vejce** . Později , po otevření ,
stromů , jen vzácně na zemi . **Holubí vejce** jsou bílá nebo nahnědlá .
spečenou žluč v kámen velikosti **holubího vejce** a břich na píď obalený tukem

(...)

3. Поиск всех прилагательных (A) в краткой форме (C), мужского рода (Y), единственного числа (S)

[tag="ACYS.*"]

to bylo včera , a ty jsi **schopen** jí všechno vyprdolit . Tak
společnost . Jeho vtip a šarm byl **znám** a poslední leč bez něj by
hladu nepoprali . Bořivoj byl **spokojen** . " Dobře jsme tu hospodu

Корпус современного американского английского (COCA) является самым большим корпусом английского языка, находящимся в свободном доступе по адресу http://corpus.byu.edu/coca/, и единственным большим и сбалансированным корпусом американского варианта английского языка. Он был создан М. Дэвисом (M. Davies, Brigham Young University, США) в 2008 г. В начале 2013 г. объем COCA, включающего тексты с 1990 г. по 2012 г., равномерно представляющие уст-

ную речь, художественную прозу, популярные журналы, газеты и научную литературу, составил 450 млн слов. Он обновляется два раза в год и удобен для наблюдения за текущими изменениями, происходящими в языке.

Из немецких текстовых корпусов необходимо упомянуть о корпусе ***Корпус немецкого языка*** **DeReK*o*** (das Deutsche Referenz Korpus), доступном по адресу http://www.ids-mannheim.de/kl/projekte/korpora/. Электронное собрание, созданное в рамках корпусного проекта Института немецкого языка в Мангейме (Германия), состоит из беллетристики, научных и публицистических текстов и содержит более 5 млрд словоупотреблений (по данным на март 2013 г.). Этот корпус оформлен как собрание отдельных немецкоязычных подкорпусов. Корпус содержит основанную на SGML морфосинтаксическую разметку, разработанную в соответствии с рекомендациями TEI. Корпусный менеджер COSMAS II, которым снабжен немецкий корпус, позволяет осуществлять поиск по лексическим единицам и по морфологическим признакам словоформ.

В настоящее время активно развивается технология создания корпусов на базе текстов из Интернета, позволяющая создавать «миллиардные» корпусы.

3.2.2. Корпусы русского языка

3.2.2.1. Первые корпусы русского языка

Первый русскоязычный корпус был создан в 1980-е годы в Университете Уппсалы, Швеция. Однако еще до первых русскоязычных корпусов в 1960–70-е годы был создан частотный словарь русского языка под рук. Л.Н. Засориной, построенный на основе корпуса текстов объемом в 1 млн словоупотреблений и включавший примерно в равной пропорции общественно-политические тексты, художественную литературу, научные и научно-популярные тексты из разных областей и драматургию.

В 1985 г. в СССР по инициативе академика А.П. Ершова были начаты работы по созданию Машинного Фонда русского языка. В создании фонда принимали участие более 40 организаций-соисполнителей, среди них Институт русского языка, Московский, 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 велось создание программных средств для проведения лингвистических исследований. [25]. 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., главнейшие словари русского языка, краткая академическая грамматика, созданы текстовые корпусы поэзии, художественной прозы, общественно-политических и технических текстов. Однако после 1991 г. в новых экономических условиях работы по созданию фонда постепенно стали сокращаться и наконец совсем прекратились.

Уппсальский корпус русского языка (Upsal'skij korpus russkix tekstov) состоит из 600 текстов, его объем составляет 1 млн словоупотреблений, поровну распределенных между образцами специальной и художественной литературы. По замыслу создателей, корпус должен был отражать современное состояние русского языка. Цель формирования корпуса заключалась в том, чтобы представить, в первую очередь, литературный язык, поэтому в массиве нет образцов разговорной речи.

В корпус отбирались специальные тексты (включались не фрагменты текстов, а целые тексты) с 1985 по 1989 г. и художественные тексты с 1960 по 1988 г. 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 отдано 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 lu†we u„snit' sut' i nazna†enie processov obnovleni„, blihnie i dal'nie celi preobrazovanij, opredelit' svoe* otnowenie k nim

Упсальский корпус входит в так называемые «Тюбингенские корпусы русских текстов», созданные в рамках работы специального научно-исследовательского сектора SFB 441 Тюбингенского университета в 1990–2000-е гг. с возможностью поиска online (http:// www. sfb441.uni-tuebingen.de/b1/rus/korpora.html#uppsalakorpus).

Корпусы размечены тегами морфологической аннотации. Разметка была осуществлена при помощи статистического теггера (TnT). Поиск может производиться как по словоформам, так и по морфологическим

тэгам. 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. был создан на Филологическом факультете МГУ в 2000–2002 гг. в Лаборатории общей и компьютерной лексикологии и лексикографии. 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 ежедневных и неежедневных («МН», «Новая газета»), «левых» («Завтра», Правда», «Правда-5») и «правых», центральных и местных, общих и профессионально ориентированных газет (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.

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 (с последующими контролирующими процедурами), а также морфемная сегментация словоформ и лексем.

Обобщение жанровых характеристик привело к объединению конкретных жанров в 9 жанровых типов.

1. Собственно информационные жанры, содержанием которых является информация, представленная в максимально объективной форме, лишенной авторской индивидуальности;

2. Информационно-публицистические жанры, в которых объективное изложение информации сопровождается ее субъективной интерпретацией, эмоциональной или интеллектуальной оценкой. Следует отметить, что в эту группу попали и неспецифические для газеты жанры: биография, заявление, приметы.

3. Собственно публицистические жанры, содержанием которых является переработанная автором информация: доказательство какого-либо положения, мнение, выражение чувств и т. д. Объективно новая для читателя информация играет здесь второстепенную роль.

4. Художественно-публицистические жанры, в которых используются различные приемы изобразительности, создания художественного текста.

5. Рекламные жанры, включающие как чисто рекламные тексты, так и рекламные сообщения, облеченные в форму традиционных газетных жанров (заметки, интервью).

6. Художественные жанры.

7. Разговорные жанры.

8. Официально-деловые жанры.

9. Прочие, куда включены такие развлекательные жанры, как игра, кроссворд, гороскоп и т. д., жанр религиозной проповеди, а также такие жанры, отнесение которых к определенному типу пока затруднительно.

3.2.2.2. Современные корпусы русского языка

3.2.2.2.1. Национальный корпус русского языка

Долгое время не было общедоступного, представительного и размеченного корпуса русского языка, с которым могли бы работать лингвисты. Непосредственная работа по созданию такого корпуса началась только в 2000 г., хотя определенные наработки существовали с 1980-х [32].

Национальный корпус русского языка (НКРЯ) – это информационно-справочная система, основанная на собрании русских текстов в электронной форме. Он был впервые размещен на сайте http://ruscorpora.ru/ в апреле 2004 г. Корпус предназначен для всех, кто интересуется различными вопросами, связанными с русским языком: профессиональных лингвистов, преподавателей языка, школьников и студентов, иностранцев, изучающих русский язык.

Национальный корпус русского языка отвечает критерию репрезентативности и другим требованиям, предъявляемым к современным

корпусам, о чем свидетельствуют следующие его характеристики (см. статистику http:// ruscorpora.ru/corpora-stat.html):

1) объем НКРЯ, который в сумме составляет около 500 млн словоупотреблений (по данным сайта на март 2013 г.);

2) 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);

3) чрезвычайно разнообразный по основным социологическим параметрам (возрасту, уровню образования и владения языком, профессиональной принадлежности, типам речевых культур) состав авторов, чьи произведения вошли в корпус (не менее 20 тыс.);

4) наличие в НКРЯ текстов, относящихся к разным периодам создания, что позволяет проследить изменения в употреблении языковых явлений, и, возможно, установить динамику этих изменений [11].

В корпусе можно условно выделить две части – современную и диахроническую. Корпус современных текстов составляют тексты, период создания которых укладывается в рамки 1951–2010 гг. Объем этой части корпуса – 230 млн словоупотреблений. Диахроническая часть объединяет тексты XVIII, XIX и 1-й половины XX вв.

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 акцентологических вариантов и изменений в акцентной системе русского языка;

5) исследование лексической вариативности, в частности, изменения состава синонимических рядов и тематических групп, а также семантических соотношений в них [11].

Национальный корпус русского языка в настоящее время помимо основного корпуса включает следующие подкорпусы:

- газетный корпус, охватывающий статьи из средств массовой информации 2000-х гг.;
- глубоко аннотированный (синтаксический) корпус, содержащий тексты, снабженные морфо-синтаксической разметкой, где помимо морфологической информации, приписанной каждому слову текста, для каждого предложения задана его синтаксическая структура (дерево зависимостей);
- совокупность параллельных двуязычных корпусов текстов разных языков (английский, французский, немецкий, испанский, итальянский, польский, украинский, белорусский), в которых можно найти все переводы для определенного слова или словосочетания;
- корпус диалектных текстов, включающий запись диалектной речи различных регионов России с сохранением их грамматической специфики; предусмотрен специальный поиск с учетом диалектной морфологии;
- корпус поэтических текстов, содержащий стихотворные произведения от XVIII в. до современности, в котором возможен поиск не только по лексическим и грамматическим, но и по специфическим для стиха признакам (поиск определенного сочетания в сонетах, в эпиграммах, в стихотворениях, написанных амфибрахием, с определенным типом рифмовки и т. п.);
- акцентологический корпус (корпус истории русского ударения), включающий тексты, несущие информацию об истории русского ударения; реализован поиск по месту ударения и просодической структуре слова.
- обучающий корпус русского языка – корпус со снятой омонимией, разметка которого ориентирована на школьную программу русского языка;
- корпус устной речи включает расшифровки магнитофонных записей публичной и частной устной речи, а также транскрипты кинофильмов 1930–2000-х гг. [27];
- мультимедийный русский корпус (МУРКО), образованный фрагментами кинофильмов 1930–2000-х гг., представленных в виде параллельных видеоряда, аудиоряда и текстовой расшифровки звучащей речи, а также 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.

3.2.2.2.2. Хельсинский аннотированный корпус (ХАНКО)

Корпус создан в Хельсинском университете в начале 2000-х гг. как часть проекта «Функциональный сиснтаксис русского языка» (рук. проф. А. Мустайоки) и постоянно развивается. Объем корпуса – 100 тыс. словоформ. Доступен в Интернете по адресу: http:// http:// www.ling.helsinki.fi/projects/hanco/. Объем корпуса – 100 тыс. словоформ. Корпус создан на основе статей из журнала «Итоги» за 2001 г. В корпусе реализованы морфологическая и синтаксическая разметки и, соответственно, морфологический и семантический поиск. Особенность корпуса – тщательно проработанный формат лингвистического описания данных и полная визуальная проверка результатов автоматической разметки, имеющая следствием полное снятие грамматической омонимии. Синтаксическая разметка корпуса должна совмещать разметку в терминах членов предложения (уже реализована) и в терминах деревьев зависимостей. Кроме того, размечены многословные устойчивые обороты (примерно 2000 единиц). Планируется семантическая разметка в терминах семантических категорий.

3.2.2.2.3. Корпусы университета г. Лидс

В 2000-е 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 корпусы и русского языка (http://corpus.leeds.ac.uk/ruscorpora.html). Это версия Национального корпуса русского языка объемом в 116 млн словоупотреблений (на основе ее был создан Частотный словарь русского языка (http://dict.ruslang.ru/freq.php). Кроме того, на этом сайте представлены Корпус русских газет (2001–2004 гг., 76 млн словоупотреблений), корпус русских текстов из Интернета (160 млн словоупотреблений), корпус деловой и экономической информации (12 млн словоупотреблений), и объединенный корпус, составленный из всех вышеназванных.

Поисковый интерфейс Leeds CQP базируется на корпусном менеджере IMS Corpus Workbench и предоставляет интересные возможности. Он позволяет вести очень точный лексико-грамматический поиск, поскольку дает возможность использовать специальный язык запросов, в том числе и с использованием языка регулярных выражений. Имеют-

ся способы управления выходным интерфейсом, формой представления результатов поиска. Можно также получить списки коллокаций, вычисленных и упорядоченных на основе ассоциативных мер MI, T-score, Log-likelihood. Там же имеется коллекция различных программных средств для обработки корпусных текстовых данных (http://corpus.leeds.ac.uk/tools/).

3.2.2.2.4. Другие текстовые корпусы русского языка

Корпус Библиотеки Мошкова. На сайте группы АОТ (http://aot.ru/search1.html) имеется большой корпус русских текстов объемом 680 млн слов, созданный А. 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. Там же имеется сервис поиска биграмм (54 млн), вычисленных по мере MI.

Корпусы в системе Sketch Engine. Английская лингвистическая служба Lexical Computing Ltd. (A. Kilgarriff) предоставляет на коммерческой основе доступ более чем к 40 корпусам различных языков. Среди них имеется ряд корпусов русского языка и прежде всего корпус, созданный из текстов Интернета по технологии Wacky объемом 20 млрд словоупотреблений. Английские исследователи совместно с чешскими разработчиками из Университета им. Масарика разработали корпусный менеджер Sketch Engine (http://sketchengine.co.uk/).

Менеджер обладает многими уникальными возможностями. Помимо стандартного поиска с выдачей конкорданса он выдает списки коллокаций по отдельным синтаксическим моделям, формирует частотный словарь, группирует лексические единицы в лексико-семантические поля с внутренней кластеризацией и указанием силы связи между лексемами.

3.2.2.2.5. Устные корпусы русского языка

Отдельно остановимся на новом типе корпуса, который, насколько нам известно, отсутствует в других национальных корпусах. В настоящее время ведется активная работа по созданию мультимедийных (в другой терминологии – мультимодальных) корпусов русского языка. Мультимедийный корпус – это электронный ресурс, предназначенный для изучения звучащей речи, «погруженной» в обстоятельства ее произнесения. Корпус такого рода, кроме текстовой составляющей, может включать видео- и аудиозаписи процесса коммуникации с привязкой к тексту. Тексты выравнивают с их расшифровками, что позволяет ис-

следовать не только языковые единицы, но и речевые действия говорящего в различных ситуациях общения, его неречевое поведение (мимику, жесты, позы). Мультимедийный русский корпус (МУРКО) был открыт для общего доступа в конце 2010 г. Он включает как кинематографические тексты, так и некинематографический материал. Помимо стандартной разметки НКРЯ (морфологическая, семантическая, метатекстовая), стандартной разметки устных текстов (социологическая, акцентологическая) и стандартной разметки МУРКО (орфоэпическая, разметка вокалической структуры), авторами проведена разметка глубоко аннотированного МУРКО (разметка речевых актов, повторов, междометий и вокальных жестов, манеры говорения и др., разметка жестов) [11].

Мультимедийные корпусы являются перспективными с точки зрения исследования взаимодействия вербальной и невербальной составляющей естественного диалога. Поскольку устная речь, а именно, непубличная устная импровизированная речь, по мнению многих ученых, является важнейшей разновидностью языка, располагающейся ближе всего к его «ядру» и демонстрирующей наиболее характерные образцы речи [57], необходимо остановиться на возможностях использования корпусов устной русской речи.

Так, задача одного из исследований с применением мультимедийного корпуса заключалась в том, чтобы показать, какие отдельные признаки жестов-иллюстраторов указывают на наличие границ сегментов дискурса [28]. Для целей исследования был создан *Корпус устных рассказов* на русском языке, стимулом для которых послужил 6-минутный видеосюжет, так называемый «Фильм о грушах» (“Pear film”). Об этом фильме было записано 8 рассказов, сделанных студентами МГУ, общей продолжительностью около 20 минут. 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 [22].

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 [28].

Мультимодальные корпусы включают видеозапись участников коммуникации, поэтому с их помощью можно исследовать эмоции. *Русскоязычный эмоциональный корпус* (REC), размеченный с учетом данных о мимике, движениях рук, бровей и т. д., позволяет изучить стратегии эмоционального взаимодействия и конфликта, непрерывное коммуникативное поведение, хезитации и речевые сбои и др. Он может также использоваться как материал для обучения работников клиентских служб или как база данных эмоциональных реакций для мультипликаторов и режиссеров [91].

В Иркутском государственном лингвистическом университете идет работа по созданию *Учебного Мультимодального Корпуса* (УМКО) видеозаписей неподготовленных учебных диалогов носителей и «не носителей» русского и китайского языков по определенным темам, размеченных в программе ELAN и представленных также в виде параллельных корпусов, выровненных по смысловым блокам внутри диалогов. Например, диалог носителей русского языка на русском языке сопоставляется с диалогом на ту же тему («Знакомство», «Регистрация в аэропорту» и др.) китайцев, изучающих русский язык, на русском языке. 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.

В качестве еще одного примера рассмотрим речевой разрабатываемый в Институте филологических исследований Филологического факультета СПбГУ. *Один речевой день* (ОРД) – звуковой корпус современного русского языка повседневного общения. 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 [35]. Помимо речи информантов, в корпусе представлены записи их коммуникан-

тов (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 45 часов (по данным на март 2013 г.).

Данный звуковой корпус позволяет изучать лингвистическую динамику записанного материала: исследовать временны*е* 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, а также решает множество других задач, таких как анализ влияния профессии на бытовую жизнь человека, получение информации о среднем артикуляционном темпе спонтанной речи носителей русского языка [38, 35].

3.2.2.6. Специальные корпусы

Специальный корпус текстов – это сбалансированный корпус, как правило, небольшой по размеру, подчиненный определенной исследовательской задаче и предназначенный для использования преимущественно в целях, соответствующих замыслу составителя.

Примером может быть *Санкт-Петербургский учебный корпус текстов школьников, изучающих английский язык (*SPbEFLLC), созданный на кафедре прикладной лингвистики РГПУ им. А.И. Герцена. 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 не-носителей языка. В репертуаре грамматических структур, обнаруженных в SPbEFLLC, есть такие, которые представляют собой случаи «переходной грамматики» (интеръязыка), выражающиеся, например, в наруше-

нии 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 интерязыка [18].

Сложным объектом с точки зрения создания и стандартизации являются исторические корпусы, такие как *Санкт-Петербургский Корпус Агиографических*[1] *Текстов* XV–XVII *вв.* (СКАТ), доступный на сайте http://project.phil.pu.ru/skat. СКАТ – это электронный корпус текстов по памятникам древнерусской агиографической литературы, 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 [1]. В 2011 г. объем корпуса составил 500 тыс. словоупотреблений [35].

Специальные корпусы текстов могут быть востребованы не менее, чем национальные. Любой *отраслевой* специальный корпус текстов может пригодиться и в данной конкретной отрасли, и в смежных областях (кораблестроение, металлы, экология, навигация и т. д.), поскольку он дает специалисту самое главное – термины в их профессиональном конкретном окружении (что тот или иной автор имеет в виду под данным термином, какое понятие за ним стоит), позволяет отследить изменения в терминологии, включая появление новых терминов.

В числе других специальных корпусов следует упомянуть Регенсбургский диахронический корпус русского языка (древнерусские тексты), корпус рукописных памятников Древней Руси (берестяные грамоты, летописи, рукописные книги), параллельный корпус пере-

---

[1] Агиогра́фия (от греч. ἅγιος «святой» и γράφω «пишу») – научная дисциплина, занимающаяся изучением житий святых, богословскими и историко-церковными аспектами святости. Жития святых могут изучаться с лингвистической, историко-богословской, исторической, социально-культурной и литературной точек зрения.

водов «Слова о полку Игореве», корпус русского электронного наследия «Манускрипт».

3.3. Корпусные исследования

3.3.1. Пользователи корпусов

Пользователей корпусов, в первую очередь, лингвистов, как правило, интересует не содержание конкретных текстов, а их метатекстовая информация и примеры употребления тех или иных языковых элементов и конструкций. Первоначальные лингвистические исследования, проводившиеся с помощью корпусов, сводились к подсчету частот встречаемости различных языковых элементов. Статистические методики используются в решении сложных лингвистических задач, таких как составление словарей и грамматик, машинный перевод, распознавание и синтез речи, средства проверки орфографии и грамматики и т. д. Так, статистическими методами на материале корпуса можно определить, какие слова регулярно встречаются вместе и, таким образом, могут быть отнесены к устойчивым словосочетаниям. Устойчивые словосочетания представляют собой с семантической точки зрения неделимую смысловую единицу, что очень важно учитывать в лексикографии и системах автоматической обработки текста.

Корпусы являются богатым источником данных для исследований по лексикографии и грамматике. С исследованиями по лексикографии тесно связаны исследования в области семантики. Наблюдая окружение той или иной лингвистической единицы в корпусе, можно установить определенные семантические признаки, характеризующие данную единицу.

Лингвисты-теоретики используют корпусы в качестве экспериментальной базы для проверки гипотез и доказательства своих теорий. Прикладные лингвисты (преподаватели, переводчики и т. д.) используют компьютерные корпусы при обучении языкам и для решения своих профессиональных задач. Особый класс пользователей представляют компьютерные лингвисты: они пытаются выявить и использовать статистические и лингвистические закономерности, присутствующие в текстах, для создания компьютерных моделей языка. Другие специалисты по языку (литературоведы, редакторы) также в ряде случаев могут получить ответы на интересующие их вопросы, обратившись к корпусу. Специалисты по общественным наукам (историки, социологи) могут изучать свои объекты через язык, используя такие параметры тек-

стов, как период, автор или жанр. Литературоведы используют корпусы для стилеметрических исследований. Наконец, корпусы используются для разработки и настройки различных автоматизированных систем (машинный перевод, распознавание речи, информационный поиск).

Чем могут корпусные данные помочь теоретической лингвистике? Они не могут заменить самонаблюдение (интроспекцию) ученого, а также обеспечить суждения лингвистов о лексике и грамматике, но они дают специалистам богатый репрезентативный эмпирический материал. Корпусы в принципе дают три типа данных, которые используются в ходе лингвистических исследований: эмпирическая поддержка, информация по частотности, экстралингвистическая информация (метаинформация). Рассмотрим эти типы данных более подробно.

3.3.2. Способы использования корпусов

3.3.2.1. Эмпирическая поддержка

Многие лингвисты используют корпус как «банк примеров», т. е. пытаются найти эмпирическую поддержку для своих гипотез, принципов и правил, над которыми они работают. Примеры, конечно, могут быть придуманы или найдены случайно, но подход корпусной лингвистики обеспечивает репрезентативность и сбалансированность языкового материала, а также поисковый инструмент, который обычно дает возможность хорошей выборки в определенном корпусе. Многие считавшиеся верными на протяжении длительного времени утверждения были опровергнуты корпусными данными. Было, например, опровергнуто часто повторявшееся утверждение о том, что частицы в немецких глаголах с отделяемыми приставками не могут встречаться в начале предложения.

В корпусах текстов было найдено достаточно много грамматически правильных примеров начальной позиции частицы [49]. Подобно этому, предложения, объявленные сторонниками генеративной лингвистики грамматически неправильными, скорее, должны считаться грамматически правильными, потому что подобные структуры на самом деле встречаются в современном английском языке (как выявлено в Интернете, использованном в данном случае как веб-корпус, см. п. 3.1.4):

- Harry reminds me of himself [Postal 1970]; ср. Интернет: Joe reminds me of himself.

• John will leave until tomorrow [Lakoff 1970, с. 148]; ср. Интернет: I will leave until tomorrow.

• What an idiot I thought Tom was [Postal 1968, с. 75]; ср. Интернет: What an idiot I thought the main character to be. [19].

Об этом же пишет Н.В. Перцов, опровергающий суждения авторитетнейших лингвистов о русском языке, используя материал Национального корпуса русского языка: «...Следует признать, что возможности корпусов все-таки еще недостаточно усвоены лингвистической общественностью вообще и лингвистами в частности. Обращение к корпусным данным еще не стало столь же привычным и обязательным при формулировке и проверке тех или иных утверждений относительно фактов языка, как обращение к грамматикам и словарям, к работам коллег» [29, с. 318]. Даже в тех работах, авторы которых широко используют корпусы, часто встречаются утверждения о фактах языка, которые противоречат корпусным данным:

• *длинные глаза,* **но** *продолговатые глаза* [34, с. 71]: ср. НКРЯ: *Глеб вздрогнул: его* ***длинные глаза*** *какое-то время словно проверяли что-то во мне, ранее подвергавшееся сомнению* (Е. Маркова, 1990–2000).

• *тонкие колени,* **но** *острые колени* [34, с. 71]: ср. НКРЯ: <...> ***тонкими коленями*** *обхватила бочонок с натянутой на него пергаментно сухой кожей* <.. > (Д. Рубина, 2003).

• «<...> слово *счастье* не может обозначать ни событие (оно не может *наступить, произойти, случиться* и т. п.), ни его переживание» [16, с. 164]: ср. НКРЯ: <...> *в России счастье, по прогнозам российского президента,* ***наступит*** *только в 2010 году* («Известия», 2001.10.30) [29].

Свидетельства из корпусов могут быть найдены для верификации гипотез на каждом языковом уровне, от звуков речи до целых разговоров и текстов. Внутри этой структуры можно повторять анализ и воспроизводить результаты, что невозможно в ходе самонаблюдения.

3.3.2.2. Статистическая информация

Эмпирическая поддержка представляет собой качественный метод использования корпуса, но корпусы также подкрепляют ее информацией по частотности для слов, фраз и конструкций, которая может быть использована для разнообразных исследований. Количественные исследования (которые, конечно, часто основываются на качественном анализе) используются во многих сферах теоретической и компьютер-

ной лингвистики. Они показывают сходства и различия между разными группами говорящих или между разными типами текстов, обеспечивают данные о частотности лексических единиц и конструкций для психолингвистических исследований и т. д.

3.3.2.3. Метаинформация

В дополнение к лингвистическому контексту корпус представляет экстралингвистическую информацию, или метаинформацию, по таким факторам, как возраст или пол говорящего/ пишущего, жанр текста, временна́я или пространственная информация о происхождении текста и т. д. Она позволяет сравнивать разные типы текстов или разные группы говорящих.

По мнению многих ученых, корпусная лингвистика – не отдельная парадигма лингвистики, а, скорее, ее методология. В частности, многие известные корпусы английского языка создавались и применялись для специальных исследований представителями различных направлений лингвистики.

Так, корпус CHILDES, содержащий транскрипты детской устной речи в различных коммуникативных ситуациях, широко используется в области психолингвистики учеными, которые интересуются тем, как дети овладевают языком [52].

Хельсинкский корпус английского языка содержит различные типы письменных текстов, начиная с ранних периодов английского языка, и используется в области истории языка для изучения его эволюции [46].

Бергенский корпус английского языка лондонских подростков COLT (The Bergen Corpus of London Teenage Language) содержит речь лондонских подростков (13-17 лет) и используется в области социолингвистики для исследования языка определенной возрастной группы [58]. Лингвистов, использующих корпусы в своих исследованиях, объединяет уверенность в том, что лингвистический анализ на материале «реального» языка является предпочтительным, так как он обеспечивает более надежные результаты [53].

3.3.3. Лексикографические исследования, основанные на корпусах

Лексикографические исследования необходимы, в первую очередь, для составления словарей, а также для нужд дескриптивной и прикладной лингвистики. Перед исследованием необходимо выявить информационную потребность лексикографов. Например, основные

типы запросов автора толкового академического словаря русского языка заключаются в необходимости найти следующее:

- новое слово по времени его появления,
- исходную форму слова;
- цитаты к уже известным значениям;
- цитаты к тем значениям, которые в словаре не проиллюстрированы цитатами (чаще всего это грамматически обусловленные значения, например, страдательные формы русских глаголов или речевые употребления);
- дополнительные новые цитаты к тому или иному значению;
- новые типы лексической и синтаксической сочетаемости;
- новые фразеологизмы;
- новые современные научные толкования специальных терминов [9].

Грамматические и лексикографические модели системно взаимодействуют. В то время как традиционные подходы могут определить группу синонимичных слов, лексикографические исследования на базе корпусного подхода пытаются показать, как соотносимые слова используются в разных ситуациях и как они применяются в разных контекстах. В частности, языковые исследования, основанные на эмпирических данных, проводятся лексикографами оксфордских словарей английского языка. Для того чтобы проследить эволюцию языка, они используют Оксфордский корпус английского языка (Oxford English Corpus). Уже весной 2010 г. корпус включал более 2 млрд словоупотреблений из текстов XXI в., относящихся к разным регистрам, включая неформальные – электронные сообщения и блоги. Изменения в употреблении слов, их орфографии начинаются прежде всего в текстах подобного типа.

В учебнике Д. Байбера, С. Конрад, Р. Реппен «Corpus Linguistics. Investigating language structure and use» [40] (далее – Corpus Linguistics) выделяется шесть основных вопросов, стоящих перед исследователями-лексикографами, действующими на основе корпусного подхода:

1. Какие значения ассоциируются с конкретным словом?
2. Какова частотность слова относительно других близких к нему слов?
3. Какие нелингвистические модели имеет данное слово (по отношению к регистрам, историческим периодам, диалектам и т. д.)?
4. Какие слова обычно встречаются вместе с данным словом и каково распределение этих сочетаемостных последовательностей в разных регистрах?
5. Как распределены смыслы и типы использования слова?

6. Как используются и по-разному распределяются слова, кажущиеся синонимичными? [40]

Одно из преимуществ корпусного исследования в лексикографии состоит в том, что корпус можно использовать для демонстрации множества контекстов, в которых употребляется слово. Затем из этих контекстов, можно выделить разные смыслы, ассоциируемые со словом.

3.3.3.1. Пример одного лексикографического исследования

Одно из предназначений корпуса заключается в том, чтобы экономить усилия исследователя при изучении лексики проблемной области. В частности, корпус должен быть не просто строгим подмножеством текстов проблемной области, но, по возможности, существенно отличаться от нее по объему. В общем случае, чем более «экономичен» корпус, тем выше порог отображения. Конкордансы могут представлять слишком большое количество данных. Объем конкордансов не только для служебных, но иногда и для знаменательных слов в больших корпусах может достигать нескольких тысяч страниц, и на один интересный пример может приходиться сотня тривиальных [2].

Например, для слова deal из восьмимиллионного подкорпуса *корпуса Лонгман-Ланкастер* выдается более 1500 употреблений, что усложняет задачу сгруппировать разные смыслы слова или рассортировать их по важности. В таком случае приходится использовать дополнительные инструменты. Так, большинство программ конкордансов может создавать список частоты слов, который обычно представляется в алфавитном порядке, по порядку встречаемости или по частоте. Широкую популярность приобрела система Sketch Engine, которая выдает ограниченный набор статистических словосочетаний (коллокаций), упорядоченный по структурно-синтаксическим моделям.

Приведем из учебника Corpus Linguistics [40] пример выявления значений, ассоциируемых со словом deal в английском языке.

Анализ значения слов осложняется тем, что многие словоформы в английском языке имеют множество грамматических функций. Так, словоформа deals может быть использована как глагол в 3-м лице единственного числа и как существительное во множественном числе. Deal и dealing могут быть использованы как глагол и как существительное. Частотные списки, построенные на данных неаннотированных корпусов, ограничены в своей полезности, поскольку они не показывают, какие грамматические употребления слов являются частыми, а какие – редкими.

Для того чтобы определить, сколько раз словоформа deal встречается как существительное и сколько раз – как глагол, нужно посмотреть на формы в контексте, определить их грамматические категории и только потом осуществлять подсчет. Такое решение будет очень затратным по времени для 182 случаев встречаемости словоформы deal в LOB корпусе и тем более в других, больших по объему корпусах, и для очень распространенных слов, таких как look, которое встречается около 500 раз на 1 млн слов. Более правильное решение в таком случае – это использование аннотированного корпуса, в котором каждое слово помечено своей грамматической категорией. В таком корпусе можно произвести автоматические подсчеты для каждой грамматической формы слова отдельно.

В табл. 6 показан частотный список, выданный программой TACT (Text-Analysis Computing Tools), который показывает распределение грамматических форм слова deal в аннотированном *корпусе Ланкастер-Осло-Берген*. Грамматическая категория каждого слова следует непосредственно за словом после символа «подчерк». Так, слово deal встречается как существительное в единственном числе (граммема nn) 115 раз, как имя собственное (np) – 1 раз, как глагол (vb) – 66 раз и т. д. С такой информацией из аннотированного корпуса можно продолжать изучение встречаемости deal более подробно, обращая внимание на распределение его глагольных и субстантивных форм и сравнивая их использование в разных регистрах.

Таблица 6. Частота форм слова deal в аннотированном корпусе Ланкастер-Осло-Берген (LOB)

deal_nn	115
deal_np	1
deals_nns	5
deal_vb	66
dealing_vbg	51
deals_vbz	20
dealt_vbd	14
dealt_vbn	17

Распределение deal *по регистрам*

Слова часто употребляются по-разному в разных регистрах, поэтому всеобъемлющие характеристики слова могут не отражать реальное положение дел в языке. Сначала рассмотрим лексему deal как су-

ществительное в аннотированном корпусе Ланкастер-Осло-Берген, обращая внимание только на формы единственного и множественного числа (deal и deals).

Поскольку корпус Ланкастер-Осло-Берген составлен из текстов разных регистров, таких как научная литература, художественная проза, приключенческая литература и ковбойские романы, есть возможность сравнить частоты deal и deals в разных регистрах (табл. 7).

Табл. 7 включает абсолютные частоты (сырые) подсчеты (raw counts) и нормированные частоты (normed counts) в пересчете на 100 тыс. словоупотреблений.

Таблица 7. Частотность существительного deal в определенных регистрах, нормированная на 100 тыс. слов

Частота / Регистр	Примерное количество слов в подкорпусе	Абсолютная частота для deal	Нормированная частота для deal (на 100 тыс. слов)
Репортажи прессы	88000	14	15,9
Обзоры прессы	34000	4	11,8
Передовицы	54000	4	7,4
Религиозная литература	34000	5	14,7
Научная литература	160000	16	10,0
Научно-популярная литература	88000	11	12,5
Беллетристика	154000	24	15,6
Художественная проза	58000	5	8,6

Подкорпусы регистров включают различное количество слов: в репортажах прессы – 88 тыс. слов, в обзорах прессы – 34 тыс. слов, а в научной литературе – 160 тыс. слов и т. д. По этой причине абсолютные показатели нельзя использовать как критерий для вывода о большей или меньшей частотности слова в одном регистре по сравнению с другим. Так, в репортажах прессы анализируемое слово встретилось 14 раз, а в научной литературе 16, но это ни о чем не говорит. Поэтому для сравнений используют «нормированную» (нормализованную) частоту. ***Нормированные*** частоты получаются преобразованием количества случаев встречаемости слова по стандартной шкале, обычно в пересчете на 1 млн слов или, в данном случае, на 100 тыс. слов. Когда

подсчеты нормированы, в репортажах прессы получается 15,9 случаев встречаемости на 100 тыс. слов, а в научной литературе – всего 10 случаев на 100 тыс. слов и т. д. Следовательно, только нормированные подсчеты обеспечивают достоверные основания для сравнения по регистрам.

Когда случаи встречаемости существительного deal распределены по регистрам, проблема размера корпуса для лексикографической работы становится еще более очевидной. Табл. 7 показывает, что в четырех из восьми подкорпусов отмечено всего 4–5 случаев встречаемости. Ни в одном из регистров нет достаточно большого количества употреблений deal, максимальное количество – 24 (художественная проза). Понятно, что корпус Ланкастер-Осло-Берген слишком мал для детального анализа использования deal в качестве существительного, поэтому далее будут рассмотрены модели его распределения по регистрам на материале более солидного по объему корпуса Лонгман-Ланкастер.

Таблица 8. Частотность существительного и глагола deal в подкорпусах из двух регистров корпуса Лонгман-Ланкастер

Частота / Регистр	Примерное количество слов в подкорпусе	Нормированная частота (на 1 млн слов)	
		Существительное	Глагол
Всего	4000000	90	119
Художественная проза	2000000	107	63
Научная литература	2000000	74	176

Табл. 8 показывает, что в корпусе Лонгман-Ланкастер deal и deals встречаются намного чаще, и это обеспечивает более солидную базу для анализа их употребления. Благодаря данным этой таблицы частотности становятся очевидными несколько интересных моделей.

Во-первых, нормированные подсчеты для всех примеров текстов показывают, что deal/ deals как глагол лишь ненамного чаще встречается, чем deal/ deals как существительное (119 слов на млн словоупотреблений в сравнении с 90 словами на млн). Однако если рассмотреть встречаемость по регистрам, появится другая картина. В научной литературе deal/ deals функционирует как глагол в два раза чаще, чем как существительное (176 против 74 на млн слов). Художественная проза показывает противоположную модель, в которой употребление deal/

deals в качестве существительного намного чаще, чем в качестве глагола (107 против 63 на млн слов).

Эти модели употребления deal высвечивают и другой важный момент в создании корпуса: корпус, ограниченный одним из регистров, не будет представлять язык в других регистрах. Так, невозможно сделать обобщения на материале одного регистра для моделей других регистров. Пример показывает, что относительная частота deal как существительного и как глагола в научной литературе является полностью противоположной их относительной частоте в художественной прозе. Корпус, ограниченный любым из этих регистров, совсем не показал бы того, что найдено в другом регистре, и построение моделей языкового использования этого слова было бы неверным.

Кроме того, этот пример показывает, какими ошибочными и недостоверными могут быть всеобъемлющие обобщения. Они скрывают противоположные модели использования, которые в действительности имеют место, и в результате часто являются неточными для любой разновидности, описывая тип языка, который вообще-то в действительности не существует. Чтобы ответить на вопрос, чем можно объяснить разное распределение субстантивных и глагольных форм по регистрам, нужно проанализировать разные смыслы слова и способы его употребления в каждом регистре.

Распределение смыслов (значений) по регистрам

Корпусы позволяют исследовать значения слов путем использования конкордансов. Начать исследование смыслов слов можно с анализа их ***коллокатов*** (collocates) – слов, с которыми анализируемое слово (часто) встречается вместе. Для каждой коллокации (collocation) существует сильная тенденция ассоциироваться с одним смыслом или значением (хотя более чем одно сочетание может ассоциироваться с тем же смыслом). Поэтому, выделяя наиболее частые коллокации слова, можно эффективно и надежно анализировать смыслы. Далее нужно сравнить то, что демонстрирует анализ коллокатов существительного deal в корпусе, с его словарными дефинициями.

В табл. 9 приведены коллокаты для существительного deal в двух регистрах из корпуса Лонгман-Ланкастер. Подобные таблицы, показывающие список сочетаемости, отсортированный по частоте, можно получить с помощью различных программ конкордансов.

Левые коллокаты – это слова, которые непосредственно предшествуют существительному deal. Например, из данных, представленных в табл. 9, следует, что слово good является частым левым коллокатом

для deal. Правые коллокаты – это слова, которые непосредственно следуют за существительным deal. Например, слово of является частым правым коллокатом для deal. Списки в этой таблице представляют только первое слово вправо и влево от deal, но те же технологии позволяют исследовать сочетаемость на расстоянии (например, на расстоянии двух или трех слов). Как видно из таблицы, в научной литературе самым частым левым коллокатом существительного deal является прилагательное great (45 раз на млн слов), затем следует прилагательное good (23 раза на млн слов).

Таблица 9. Частотные коллокаты существительного deal в двух подкорпусах корпуса Лонгман-Ланкастер (5,7 млн слов)

Подкорпус	Нормированная частота (на 1 млн слов)
Научная литература *(подкорпус 2,7 млн слов)*	
Левые коллокаты	
great	45
good	23
Правые коллокаты	
of	39
more	7
in	3
to	3
Художественная проза *(подкорпус 3 млн слов)*	
Левые коллокаты	
great	40
good	28
the	8
big	3
Правые коллокаты	
of	28
to	7
about	5
more	3
with	3

Следующие по порядку коллокаты (package и that) встретились дважды, и поэтому не были внесены в табл. 9, которая показывает

только коллокаты, встретившиеся хотя бы 3 раза на млн слов в каждом регистре.

В научной литературе, очевидно, коллокации good deal и great deal будут обозначать большое количество чего-либо или операции в бизнесе. При рассмотрении правых коллокатов становится понятным, что значение, относящееся к количеству, является для deal наиболее частотным. Это подтверждает и частотность правого коллоката of (39 раз на млн слов). Частотность следующего коллоката намного меньше, ср.: more (7 раз на млн слов), in и to (3 раза на млн слов). Итак, существительное deal чаще всего имеет значение количества, как в словосочетаниях a good/great deal of.

Анализ смыслов слова можно проверить, просмотрев списки конкордансов, которые показывают данные словосочетания. Например, частые употребления a good/great deal of включают a good deal of work и a good deal of attention. Другие наиболее частые правые коллокаты также имеют отношение к количеству. Например, коллокат more используется в словосочетаниях a great deal more tolerance и a good deal more inhabited. Коллокаты in и to тоже используются с существительным deal для обозначения количества, ср.: a great deal in common, differ a great deal in their understanding, a great deal to be desired, a great deal to offer.

Табл. 9 показывает, что словосочетания в художественной прозе имеют интересные сходства и различия со словосочетаниями в научной литературе. Снова самыми частыми левыми коллокатами являются great и good. Действительно, совместная встречаемость good deal и great deal очень похожа в двух регистрах (примерно 68 раз на млн слов). Однако в художественной литературе есть существенное количество случаев встречаемости (96 примеров), не относящихся к модели «good/great + deal»: the встречается 8 раз на млн слов и big встречается 3 раза на 1 млн слов. Остальные 37 коллокатов встречаются не чаще двух раз.

Модель словосочетаний предполагает, что значение количества является для художественной прозы центральным, но не единственным. Например, левый коллокат the используется с deal в значении договоренности, как в примерах part of the deal *is*... и Isn't that the deal? Коллокат big представляет другой смысл. Он показывает отсутствие важности в выражениях типа no big deal и what's the big deal?

Кроме того, многие коллокаты, не самые частотные, ассоциируются с операциями в бизнесе: property deal, record deal, cash deal, land

deal, mining deal. Хотя эти слова не относятся к пяти основным коллокатам deal, вместе они демонстрируют важный смысл.

Словосочетания с существительным deal в художественной прозе также раскрывают значение, не найденное в научной литературе. В списке правых коллокатов есть 4 случая встречаемости table и 1 случай встречаемости box, когда речь идет о типе дерева (древесины). Эти коллокаты не являются частотными, но они указывают на еще одно использование deal в художественной прозе, а их встречаемость в 5 разных текстах говорит о том, что это использование относительно распространено.

Далее в учебнике Corpus Linguistics сравниваются полученные на основе корпусного подхода частоты со словарными определениями существительного deal. Обзор словарей показывает поразительное разнообразие его значений. Некоторые словари дают только одну главную статью, другие – целых 4. В словарных статьях количество определений варьирует от 2–3 до 20–30. При таком разнообразии представления пользователю достаточно сложно догадаться, каковы наиболее частые значения существительного deal.

Табл. 10 показывает 7 значений существительного deal, которые наиболее часто повторяются в пяти словарях. Большинство словарей упоминает все 7 значений, однако порядок их расположения различен. Например, значение «большое, но неопределенное количество» вводится в первой словарной статье в Webster's Third Dictionary в дефиниции 2 и в Random House Dictionary в дефиниции 21.

Сравнивая эти словарные определения с результатами исследования существительного deal с помощью корпусов, можно выделить несколько проблем. Употребление существительного deal в значении количества, бесспорно, является наиболее частотным для обоих анализируемых регистров корпуса. Тем не менее, этот смысл не раскрывается до 16-й или даже 23-й дефиниции в двух словарях.

Кроме того, анализ коллокатов обнаружил относительно частотное значение, не отмеченное в этих словарях: использование big deal в значении «незначительности». Наконец, во всех пяти словарях регистровые отличия не принимаются во внимание, хотя более поздние словари, созданные на основе корпусных данных, начинают учитывать важные регистровые модели.

Таблица 10. Словарные дефиниции существительного deal

Словари / Значение	Webster's Encyclo. 1989	Webster's Third 1981	Chamers 1993	Random House 1993	Longman Lang. and Culture 1992
large but indefinite amount	entry 1 sense 13	entry 3 sense 3	entry 1 sense 2	entry 1 sense 3	entry 1 sense 21
agreement/arra ngement	entry 1 sense 16	entry 1 sense 16	—	entry 1 sense 18	entry 2 sense 1
distribution of cards in a game	entry 1 sense 18	entry 1 sense 3	entry 1 sense 4	entry 1 sense 21	entry 2 sense 4
Treatment received	entry 1 sense 15	entry 3 sense 2	entry 1 sense 6	entry 1 sense 6	entry 2 sense 2
act of distributing	entry 1 sense 17	—	—	entry 1 sense 23	—
pine or fir wood	entry 2 3 senses	entry 4 2 senses	entry 2 sense 1	entry 2 3 senses	entry 3 sense 1
act of buying or selling a business transaction	entry 1 sense 13	entry 3 sense 2	entry 1 sense 5	entry 1 sense 17	entry 2 sense 1

Однако здесь следует подчеркнуть один очень важный момент, называемый «отступлением» корпусной лингвистики. Корпусная лингвистика не отрицает ценности и необходимости речевых данных, не представленных в корпусной форме, и признает, что из корпуса текстов нельзя извлечь все возможные лингвистические выводы, т. е. то, что корпус текстов не является самодостаточным [31]. Все пять словарей указывают не обнаруженное в ходе корпусного исследования значение, относящееся к раздаче карт в игре. Хотя это одно из первых значений, которые говорящие ассоциируют с существительным deal, употребляется оно не часто (кроме карточных игр!). Этот пробел высвечивает важность больших представительных корпусов для лексикографической работы. Он также показывает, как основанный на корпусе анализ нуждается в проверке интуицией носителя языка. Словарь должен включать значение существительного deal в карточной игре, даже если оно ни разу не встретилось в корпусе – каждый носитель английского языка узнает его. Однако важно полагаться и на корпусный ана-

лиз, который говорит, что это одно из относительно редких употреблений существительного deal, которое вряд ли встретится изучающим английский язык, помимо ограниченных областей использования. Таким образом, лексикографическая работа должна объединять обе перспективы: выделять все значения, но указывать наиболее частые или важные значения, принимая во внимание их регистровую отнесенность.

Слово deal *как глагол*

Значения deal как глагола в учебнике Corpus Linguistics рассматриваются на словосочетаниях с использованием той же технологии. Коллокация deal with является примером того, как пара сочетающихся слов может ассоциироваться с разными смыслами.

Пара deal with встречается намного чаще, чем другие коллокации как в научной литературе, так и в художественной прозе. В *корпусе* Лонгман-Ланкастер эта пара встречается примерно 157 раз на 1 млн слов в научной литературе и 58 раз на 1 млн слов в художественной прозе. Для сравнения, следующий наиболее частый правый коллокат глагола deal в научной литературе – это only (2,6 случаев встречаемости на 1 млн слов). В художественной прозе следующая коллокация – deal in (4,6 случаев встречаемости на 1 млн слов).

Конкордансы для коллокации deal with наглядно демонстрируют несколько разных смыслов, наиболее частый из них – это «то, о чем пойдет речь в книге, статье, исследовании и т. д.». Просмотр всего списка конкордансов показывает, что это значение намного чаще встречается в научной литературе, и в этом регистре коллокация deal with имеет бóльшую частотность: The second controversy dealt with the source of nitrogen in plants; An important point to note is that the preceding discussion has dealt with thermodynamic acidity...; Other environmental effects are dealt with in other chapters.

Еще одно значение коллокации deal with – «решить проблему»: When they had dealt with the fire another crisis arose; Moreover many losses are due to chilling and crushing, both factors that can be dealt with by good environmental control and housing.

Преимущественно в художественной прозе появляется также значение «справляться с ситуацией» каким-либо способом без действительного решения проблемы: He didn’t have the right temperament to deal with the Hennigs of this world; She would have rather there been a fight, anger – or even tears and pleadings. These she could deal with, not

this deadly coldness exhibited by Alice; But suffering is also a fact. It has to be dealt with.

Наконец, коллокация deal with имеет значение, обозначающее взаимодействие с человеком, особенно с партнером по бизнесу: Hansie De Beer runs the farm, he's the one Mehring usually deals with; The son handed me a small suitcase with the distant eyes of a man dealing with a chauffeur.

Очевидно, что эти значения коллокации deal with заслуживают более серьезного анализа. Корпус можно проанализировать на предмет поиска слов, встречающихся после deal with, закодировав их по разным семантическим категориям (предмет разговора, проблемы и т. д.). В таком случае, расширенные сочетаемостные рамки можно будет использовать для разграничения этих смыслов. Однако здесь было важно показать, что коллокации не обязательно всегда ассоциируются с одним и тем же значением. Напротив, в некоторых случаях одна и та же коллокация может употребляться в разных значениях, что проявляется в более широком контексте. Для полноценного и полномасштабного исследования смыслов и смысловых коллокаций желательно иметь семантически размеченные корпусы.

3.3.3.2. Анализ использования слов, кажущихся синонимами

В языках есть много слов, которые считаются синонимами, словари и тезаурусы часто характеризуют их как идентичные по значению. Однако модели употребления синонимичных слов обычно сильно различаются. Лексикографический анализ, базирующийся на корпусных данных, особенно хорошо служит раскрытию таких системных различий в моделях использования.

Авторы учебника Corpus Linguistics рассматривают синонимичные английские слова big, large и great. Тезаурусы часто перечисляют слова big, large, great как синонимы размера. Ранее было показано, что great deal часто используется, когда речь идет о большом количестве чего-либо. Чтобы исследовать различные употребления great относительно big и large нужно проанализировать дополнительные сочетания.

Табл. 11 показывает частотные распределения big, large, great в 5.7-миллионном фрагменте из корпуса Лонгман-Ланкастер (Longman-Lancaster Corpus). Из таблицы видно, что, когда регистры объединены, мы получаем абсолютные частоты, которые не отражают действительное положение дел ни в одном регистре. Например, объединенный подкорпус показывает, что large является наиболее частотным из этих

трех прилагательных, за ним следует great и потом big (с нормированными частотами приблизительно 408, 393 и 230, соответственно).

Таблица 11. Частотное распределение big, large, great в подкорпусе объемом 5,7 млн с/у корпуса Лонгман-Ланкастер

Частота / Слова	Ненормированные частоты	Нормированные на 1 млн с/у частоты
Весь подкорпус (5,7 млн с/у)		
big	1319	230
large	2342	408
great	2254	393
Научная литература (2,7 млн с/у)		
big	84	31
large	1641	605
great	772	284
Художественная проза (3 млн с/у)		
big	1235	408
large	701	232
great	1482	490

В научной литературе порядок трех прилагательных тот же, но разница в частотности large и big намного больше: large используется очень часто, 605 случаев на 1 млн словоупотреблений (с/у), а big – очень редко, всего 31 случай на 1 млн словоупотреблений. Такие большие различия вовсе нельзя предсказать, глядя на объединенные подсчеты. Модели в художественной прозе еще менее предсказуемы на основе объединенных подсчетов, поскольку они почти противоположны тому, что существует в научной литературе: оба прилагательных great и big встречаются часто (с частотой 490 и 408 на 1 млн словоупотреблений), тогда как large имеет намного более низкую частоту (232 раза на 1 млн словоупотреблений).

Из табл, 11 видно, что существует огромная разница в употреблении этих (кажущихся синонимичными) слов в двух регистрах: big более чем в 10 раз чаще встречается в художественной прозе, чем в научной литературе, great более чем в 1,5 раза чаще встречается в художественной прозе. С другой стороны, large в три раза чаще встречается в научной литературе, чем в художественной прозе. Для объяснения этих

различий полезно сравнить наиболее частотные коллокаты всех трех прилагательных.

Табл. 12 и 13 показывают наиболее часто встречающиеся правые коллокаты big, large и great в научной литературе и художественной прозе (примеры взяты из корпуса Лонгман-Ланкастер). Более полный анализ потребовал бы просмотра полного списка коллокатов, но здесь в фокусе внимания оказываются только 10 первых в списке коллокатов в каждом регистре, которые встречаются чаще одного раза на миллион словоупотреблений (отсюда незаполненный столбец для big).

Таблица 12. 10 наиболее частотных правых коллокатов big, large, great в научной литературе (частота нормирована на 1 млн с/у; коллокаты с частотой менее 1 на 1 млн с/у исключены)

big		**large**		**great**	
Правый коллокат	Частота на 1 млн с/у	Правый коллокат	Частота на 1 млн с/у	Правый коллокат	Частота на 1 млн с/у
enough	2,2	number	48,3	deal	44,6
traders	1,1	numbers	31,3	importance	12,5
		scale	29,4	number	8,9
		and	28,0	majority	8,1
		enough	15,9	variety	7,0
		proportion	11,8	extent	7,0
		amounts	10,7	part	4,1
		quantities	10,3	care	3,3
		part	10,0	advantage	2,6
		extent	8,9	detail	2,6
				interest	2,6

Таблица 13. Десять наиболее частотных правых коллокатов big, large, great в художественной прозе (частота нормирована на 1 млн с/у; коллокаты с частотой менее 1 на 1 млн с/у исключены)

big		**large**		**great**	
Правый коллокат	Частота на 1 млн с/у	Правый коллокат	Частота на 1 млн с/у	Правый коллокат	Частота на 1 млн с/у
man	9,6	and	15,2	deal	40,4
enough	8,9	black	4,3	man	6,6

and	8,3	enough	3,6	burrow	5,6
black	8,3	house	3,0	big	4,6
house	7,6	room	2,7	aunt	4,3
one	7,0	white	2,7	care	4,0
toe	5,0	number	2,3	pleasure	4,0
old	4,6	for	2,3	and	3,0
red	4,3	man	2,0	relief	3,0
boy	3,6	one	2,0	black	2,7
room	3,6	in	2,0	to	2,7

В обоих регистрах коллокаты big показывают, что это прилагательное чаще всего используется в отношении физического размера. В научной литературе есть только два частотных коллоката для big. Пара big enough встречается 6 раз (2,2 на 1 млн слов), обычно по отношению к физическому размеру: small trees which are not big enough to be converted into timber.

Вторая частотная пара в научной литературе – big traders, которая встретилась только 3 раза, в одном тексте об экономическом развитии в западной Африке, в котором противопоставляются крупные и мелкие торговцы.

В художественной литературе найдено много частотных коллокатов с прилагательным big. Подавляющее большинство показывает, что big используется для описания размеров физических объектов, таких как man, house, toe, boy, room и неопределенного местоимения one. Например, The big man gave me a raking glance and grinned; she was overawed by the big house; The sitting room was a big room.

Кроме того, big употребляется с другими описательными прилагательными, такими как black, old, red. Рассмотрение списка конкордансов для этих пар показывает, что они также имеют отношение к физическому размеру объектов. Например, *"The beast had teeth," said Ralph, "and* big black *eyes."*; his big black mongrel dog; the big black saucepan.

Подобным образом, пара big enough в художественной литературе используется, чтобы показать физический размер: The cart was not really big enough, he realized; The revolver, which looked big enough to stop a

florist's van, was supposed to serve as a deterrent; *'These idiotic beds aren't* big enough f*or one person, let alone two."*

Наконец, союз and очень часто является коллокатом прилагательного big. Как уже упоминалось, служебные слова являются самыми частотными в любом корпусе. And часто встречается со всеми тремя рассматриваемыми прилагательными. Тем не менее, and, являясь всегда чрезвычайно частотным, осложняет получение важных сведений об ассоциациях в данном сочетании слов. Существует ряд статистических программ для измерения силы ассоциаций между членами сочетаемостной пары по отношению к частоте каждого слова в паре (см. 3.3.3.3).

В отличие от коллокатов big, коллокаты прилагательного large в научной литературе показывают, что оно наиболее часто используется по отношению к количеству чего-либо. Семь из десяти наиболее частотных коллокатов имеют следующее значение: large + number(s), proportion, amount(s), quantities, part, extent. Например, [.. . glutinous rices] are widely grown in Asia where a large number of varieties are recognized; There are a large number of processes grouped together under the general term weathering; A large proportion of his dependent clauses are in fact noun clauses.

Large + enough также часто употребляется для указания на количество или пропорции (29 из 43 случаев встречаемости): The ratio is large enough, however, to allow; which will always tell in a finite number of steps (which may easily be large enough to require a computer) whether or not a given element of Q[x] is irreducible.

Другие случаи встречаемости large enough относятся к физическому размеру: The pore size of the agar gel and cellulose acetate is large enough that the protein molecules are able to move freely.

Наконец, пара large scale часто встречается в научной литературе по отношению к величине различных процессов. Например, This is the type of industrial organization, according to Marx, which is most compatible with large-scale centralization; [these] can then be treated in the context of large-scale motions of lithospheric plates.

В художественной прозе наиболее часто large используется в значении физического размера, встречаясь с существительными и прилагательными house, room, man, black и white. Например, a large black saucepan; the large black barrels of a sawn-off shotgun; a large white bird; Apparently Philbrick has a large house in Canton House Terrace; It was a large room, totally silent save for the voice of one sister.

Это использование large то же, что и наиболее частое использование big в художественной прозе, и многие слова используются как правые коллокаты обоих прилагательных. Однако, как показывают частоты, large не так часто используется в художественной прозе, как big, и поэтому эти коллокаты не так часто встречаются с large, как с big. Например, big man встречается 9,6 раза на миллион, large man встречается всего 2 раза на миллион. Подобно этому, big house встречается 7,6 раза на миллион, а large house – всего 3 раза на миллион.

В художественной литературе large также имеет отношение к количеству. Сочетаемостная пара large number относительно часто встречается, а другие правые коллокаты, такие как amount, proportion и sum, хотя и редкие по отдельности, вместе образуют класс коллокатов, относящихся к количеству: A large number of people sat round a table; Ichiro was fascinated by the large amount of space in our house.

Третье прилагательное, great имеет другую модель сочетаемости. В научной литературе оно наиболее часто показывает количество. Чаще всего это значение проявляется в сочетаемостной паре great deal. Тем не менее, встречаются также пары great number, great majority, great variety, great extent и great part. Например, There is not a great deal of information on the minimum size of pore into which a root can grow; The great majority of mechanical problems give rise to matrices having distinct eigenvalues.

Это употребление great в научной литературе подобно употреблению large, хотя large никогда не употребляется с deal. Однако у great есть еще одно совершенно особенное значение, показывающее интенсивность. В научной литературе это значение встречается с правыми сочетаемостями importance, care, advantage, detail и interest: It is of great importance to control ectoparasites; The figures have to be interpreted with great care; the structure of some is now known in great detail.

В художественной прозе great в основном используется для обозначения количества, в паре great deal: He stood and drank a great deal of apple juice; Sandy was a bright young woman who seemed to know a great deal about life in Alaska. Однако сочетаемости great в художественной прозе показывают, что у него гораздо больший спектр смыслов. Например, great man встречается 6,6 раз на миллион слов по отношению к значению чего-то очень важного и очень хорошего: We approach a great man through his servants; *"...He was a* great man, and we all feel as *though we've been orphaned."*

Некоторые случаи встречаемости great care также передают значение «очень хорошего»: I promise you we will take great care of him.

В дополнение к этому, great иногда имеет отношение к физическому размеру, как в примере: a great, black bird; his great black moustache.

Сочетаемостная пара great burrow (нора) также имеет отношение к физическому размеру. Эта пара опять имеет отношение к влиянию отдельного текста на корпусные данные: все случаи встречаемости этого сочетания взяты из книги о кроликах, которые встречаются в местечке, названном 'the great burrow': In the great burrow, however, things happened differently.

В дополнение к этому, great в художественной литературе употребляется как усилитель (интенсификатор) в сочетании great big: *It's a* great big country with a continent of promise; *there're* great big gaps where they cut the wire and come, out at night; All those ones who live in those great big piecrust mock- two-door houses with His and Her Caddies parked out by the hydrangea bushes. Great также употребляется в значении усиления, как в научной литературе в сочетаниях great care, great pleasure и great relief: Jimmy found great pleasure in the society of one who had seen so much of the world;He laid the conch with great care in the grass at his feet; his fingers were numb, and he had great difficulty in undoing his collar.

Наконец, у одного сочетания есть специализированное значение фамильного родства great aunt: He was almost as old as her great aunt had been; my great-aunt has appeared unexpectedly and is carrying me off to her home in Surrey.

Даже этот краткий анализ показывает, что, несмотря на различия, выделяются определенные модели, существенные в использовании этих синонимичных слов. Big чаще всего относится к физическому размеру, large относится к количеству, great также относится к количествам, особенно в сочетании great deal, но у этого прилагательного более широкий спектр значений, от усиления до терминов родства.

Эти три предпочтительные значения у трех прилагательных помогают объяснить их частотное распределение по двум регистрам. Художественная проза содержит много физических описаний, касающихся размера объектов, людей, мест. Напротив, когда о размере говорится в научной литературе, более вероятно использование специфических измерений. Предпочтительное значение физического размера big объясняет, почему в художественной прозе это прилагательное встречается чаще, чем в научной литературе. Научная литература, напротив, больше сосредоточена на количествах, что объясняет частотность large в этом регистре. Оба регистра часто используют great для обозначения

количества (great deal). Однако в художественной прозе есть намного больше разных значений у прилагательного great по сравнению с научной литературой.

Таким образом, синонимичные прилагательные совершенно не эквивалентны по своим значениям, когда действительные модели их использования анализируются на большом эмпирическом материале. Скорее, основанный на корпусных данных анализ можно использовать для того, чтобы показать, как каждое прилагательное имеет собственные предпочтительные коллокаты, различные предпочтительные значения и различное распределение по регистрам.

Удаленные коллокаты large

Словам не обязательно примыкать друг к другу, чтобы их можно было ассоциировать друг с другом. А именно, два слова могут иметь тенденцию совместно встречаться даже если между ними расположено несколько слов. Чтобы проиллюстрировать полезность рассмотрения сочетаний на расстоянии, здесь будут описаны сочетаемостные модели во втором слове вправо от large, e.g. в последовательности large X of.

Таблица 14. Наиболее частотные коллокаты слова large на расстоянии двух слов справа (частота нормирована на 1 млн с/у)

Научная литература (подкорпус объемом 2,7 млн с/у)		**Художественная литература** (подкорпус объемом 3 млн с/у)	
Коллокат	Частота на 1 млн с/у	Коллокат	Частота на 1 млн с/у
of	167,4	of	31,1
in	19,9	and	7,9
and	16,6	in	5,6
to	14,4	eyes	4,3
the	7,7	on	3,3
are	7,0	which	3,0
with	6,6	with	2,7
on	5,5	for	2,0
is	4,1	room	2,0
open	4,1	the	2,0
small	4,1	too	2,0
that	4,1	was	2,0

Табл. 14 показывает частотные коллокаты, встречающиеся в двух словах вправо от large. Сочетаемостная ассоциация large X of является

частотной как в научной литературе, так и в художественной прозе. Действительно, эта сочетаемостная ассоциация так часто встречается, что ее уже можно рассматривать как рамку (frame), позволяющую подстановку целого спектра слов, таких как large amount of, large proportion of, large group of.

Ранее мы обнаружили, что large часто употребляется с существительными количества. Эти существительные часто используются в рамке large+ N+ of. Исследование этой расширенной модели показывает, что эти сочетаемостные последовательности выполняют две основные цели: 1) маркировка количества или измерения; 2) маркировка части большей по размеру сущности. Почти все частотные коллокаты large+ N+ of в научной литературе укладываются в эту рамку:

1) маркировка количества или измерения: a large number of, large numbers of, large amounts of, large quantities of, a large volume of, large bundles of, large masses of, a large batch of, large areas of;

2) маркировка части большей по размеру сущности: a large proportion of, a large part of, a large sample of, a large fraction of, a large segment of.

Так, в этом случае самая сильная сочетаемостная ассоциация для large встречается со служебным словом of на расстоянии. Однако исследование «внутренних» (intervening) существительных показывает, что эта рамка типично используется для маркировки количества или части большего целого.

Табл. 14 идентифицирует много интересных сочетаемостных ассоциаций. Например, модель large X eyes. Эта ассоциация особенно интересна, потому что она встречается более часто, чем сочетаемостная пара large eyes; large X eyes встречается 4,3 раза на миллион слов, тогда как large eyes – только 1,6 раза на миллион слов. Причина более частой ассоциации на расстоянии заключается в том, что существительное eyes обычно сопровождается прилагательным цвета или качества, в дополнение к дескриптору large. Как показывает данная сочетаемостная модель, эти определители встречаются в следующем порядке: large+ color/quality-adjective+ eyes, как в his large hazel eyes; large brown eyes; large black eyes; very large dark eyes; large watery eyes.

3.3.3.3. Выделение коллокаций статистическими методами

Применение корпусных методов к анализу лексической сочетаемости позволяет создавать словари нового типа, в том числе, словари устойчивых словосочетаний. Использование корпусов позволяет получать данные о совместной встречаемости лексических единиц, особен-

ностях их сочетаемости, управления и т. д. Существующие словари устойчивых словосочетаний, во-первых, охватывают далеко не полный их перечень, во-вторых, часто делают это недостаточно последовательно, поэтому возникает потребность в словаре нового типа, который можно будет назвать интегрированным словарем устойчивых словосочетаний, или словарем коллокаций, и который на самом деле будет содержать разные типы устойчивых словосочетаний.

В настоящее время в лингвистике существует несколько способов для вычисления степени связанности частей той или иной коллокации. В качестве таких статистических мер могут быть выбраны меры ассоциации (MI, t-score, log-likelihood), которые чаще всего используются при вычислении степени близости между компонентами словосочетаний в корпусе.

Для проверки применимости статистических методов для русского языка и возможности выделения коллокаций на основании указанных выше мер ассоциаций М.В. Хохловой была проведена серия экспериментов (см. работу [36]).

Исследование осуществлялось с помощью корпус-менеджера CQP (http://corpus1.leeds.ac.uk/ruscorpora.html) на базе *корпуса русских газетных текстов* (The corpus of Russian newspapers) за 2001–2004 гг. объемом 78 млн словоупотреблений, созданного в университете Лидса (Великобритания) под руководством С.А. Шарова. Материалом для исследования послужили коллокации 19 существительных, которые были отобраны по следующему принципу. Первоначально из электронного частотного словаря русского языка С.А. Шарова [37] были отобраны существительные, входящие в первую тысячу самых частотных слов. Далее по Малому академическому словарю (МАС) (Словарь русского языка 1981–1984) проверялось, имеют ли данные слова омонимы, которые могли бы искажать их частоту (например, *брак* в значениях «супружество» и «изъян»; *друг друга*, где оба элемента при лемматизации возводятся к одной лемме). Слова, имеющие омонимы, исключались из списка и не рассматривались в эксперименте. Затем список оставшихся существительных сверялся с данными в словаре коллокаций русского языка Е.Г. Борисовой [6]. В случае отсутствия словарных статей для данного слова или ограниченной информации о его сочетаемости, представленной в словаре, такие слова тоже исключались из списка. Таким образом, был получен следующий список опорных слов: *власть*, *внимание*, *возможность*, *война*, *вопрос*, *дождь*, *жизнь*, *закон*, *любовь*, *место*, *мнение*, *мысль*, *ночь*, *ответ*, *помощь*, *радость*, *слово*, *случай*, *смысл*. Ниже в табл. 15 приведены данные для

первых 10 коллокаций (из 106) с опорным словом *война*, отсортированные по значению меры MI (объем взаимной информации), где Joint – абсолютная частота данной коллокации в корпусе; Freq1 – абсолютная частота первого слова биграммы, т. е. левого коллоката для слова *война*; столбцы LL score, MI, T-score – значения мер Log-likelihood, MI и t-score для данной коллокации. Как можно увидеть, в список попали сочетания, которые, с одной стороны, являются устойчивыми, а, с другой стороны, обладают довольно высокими показателями меры MI. Исследование показало, что в диапазоне значений меры MI от 0 до 1 не были найдены словосочетания, которые можно было бы причислить к устойчивым.

Таблица 15. Значения мер ассоциации для слова *война* (левый контекст)

Коллокация	Joint	Freq1	LL score	MI	T-score
необъявленный война	9	76	30,19	11,03	3,00
междоусобный война	4	54	12,43	10,35	2,00
партизанский война	45	728	135,77	10,09	6,70
рельсовый война	6	100	18,00	10,05	2,45
победоносный война	9	174	26,31	9,84	3,00
вялотекущий война	6	142	16,92	9,54	2,45
позиционный война	5	128	13,90	9,43	2,23
холодный война	171	4747	469,90	9,31	13,06
грянуть война	14	457	37,19	9,08	3,73
финляндский война	4	148	10,37	8,90	2,00

Это позволяет сделать вывод, что сочетания, значение меры ассоциации MI которых попадает в данный интервал, оказываются статистически незначимыми. Для всех полученных сочетаний наблюдается одинаковая тенденция: чем меньше значение меры, тем больше вероятность, что эти словосочетания не зафиксированы как устойчивые в словарях русского языка. Таким образом, можно сказать, что данные

о сочетаемости, приведенные в словарях, совпадают с данными, полученными на основе мер ассоциации. Большинство коллокаций (фразем), зафиксированных в словарях, оказывается в верхней части списка, составленного на основе одной из мер ассоциации. Это говорит о том, что данные коллокации имеют высокие показатели связанности.

Важным представляется тот факт, что в результате эксперимента были выделены сочетания, не зафиксированные ни в одном из словарей. Анализ подобных сочетаний показал, что биграммы, находящиеся на самом верху списка (отсортированного по убыванию по одной из мер), с некоторой долей вероятности оказываются устойчивыми и, следовательно, могут быть внесены в словарь. В нижней части списка в подавляющем большинстве случаев оказываются свободные сочетания. Списки словосочетаний, приведенные в толковых словарях за ромбом, не могут считаться полными, хотя помещаемые туда единицы и обладают некоторой степенью устойчивости. Результаты эксперимента, с одной стороны, говорят о применимости описанных статистических мер в лексикографической практике, и, с другой стороны, указывают на известную неполноту существующих словарей.

Выявление коллокаций в специализированном корпусе может иметь большое практическое значение. Например, сравнивая данные, полученные на основе корпуса писем Н.В. Гоголя, с данными, полученными на основе общеязыковых корпусов, в ряде случаев можно увидеть существенные отличия в сочетаемости, отражающие особенности авторского словоупотребления. Таким образом, можно утверждать, что описанные выше методы и средства могут также быть эффективно использованы для изучения и создания словарей языка писателей, для выявления особенностей сочетаемости в рамках того или другого стиля или хронологического периода [36].

Поиск биграмм в большом корпусе русского языка можно осуществить на сайте http://www.aot.ru/cgi-bin/bigrams.cgi.

3.3.4. Грамматические исследования, основанные на корпусах

Изучение грамматики связано с пониманием структуры языка, включая морфологию и синтаксис. В отличие от лексикографии, грамматика не имеет долгой традиции эмпирических исследований. До недавнего времени изучению того, как носители языка на самом деле эксплуатируют грамматические ресурсы своих языков, уделялось мало внимания.

Области, обойденные вниманием в традиционных исследованиях, оказались сильной чертой основанных на корпусных данных грамматических исследований, которые могут быть применены к грамматике на уровне слова, предложения, дискурса. Здесь будет рассмотрена проблема употребления и функции морфологических характеристик путем анализа их распределения по регистрам. С помощью корпуса можно соотнести распределение морфологической характеристики с контекстами ее употребления и лучше понять функции, которые она выполняет. В учебнике Corpus Linguistics [40] пути решения этой задачи проиллюстрированы распределением номинализаций (производных существительных) по трем регистрам.

Исследование морфологической характеристики в корпусе может показать как частотность и распределение характеристики, так и различие функций отдельных вариантов. В сравнении с анализом других грамматических характеристик, основанный на корпусных данных анализ морфологических характеристик относительно прост, так как морфологические характеристики могут быть выявлены с использованием функции поиска в программах для составления конкордансов путем анализа даже неаннотированного корпуса. Большинство программ позволяют пользователю искать определенные префиксы и суффиксы, такие как un-, -ment.

3.3.4.1. Распределение и функции номинализаций

Под номинализацией (субстантивацией) в отечественном языкознании обычно понимают процесс образования абстрактного существительного от глагола, а также само существительное, образованное таким способом. В европейском языкознании это понятие шире, так как номинализацией может также быть существительное, образованное от прилагательного. Например, civilization является номинализацией, производной от глагола civilize, a kindness – номинализацией, производной от прилагательного kind.

В учебнике Corpus Linguistics проанализированы четыре продуктивных суффикса: -tion/-sion, -ness, -ment, -ity (и их формы множественного числа). Помимо автоматической обработки текстов корпуса с помощью специальных программ, потребовалась также ручная обработка для того, чтобы отобрать единицы, по форме совпадающие с поисковым шаблоном (search template), но не являющиеся номинализациями (mansion, nation, city).

Анализ номинализаций проводился в трех регистрах. Первые два – научная литература и художественная проза – представлены подкорпу-

сами корпуса Лонгман-Ланкастер. Третий регистр – устная речь – представлен корпусом Лондон-Лунд (объемом 500 тыс. словоупотреблений). Все частоты здесь нормированы на 1 млн слов текста.

Таблица 16. Частотные распределения номинализаций по трем регистрам

Регистр / Частота	Научная литература (2,7 млн с/у)	Художественная проза (3 млн с/у)	Устная речь (0,5 млн с/у)
Количество номинализаций на 1 млн с/у	44 000	11 200	11 300

Табл. 16 показывает частотные распределения для номинализаций по трем регистрам. Очевидно, что в текстах художественной прозы и устной речи отмечаются близкие частотности, а в текстах научной литературы частотность номинализаций в четыре раза больше. Можно попытаться объяснить, почему регистры имеют такие разные распределения, исследуя наиболее частые формы в контексте. Авторами учебника Corpus Linguistics разработана специальная программа, которая подсчитывает каждую индивидуальную номинализацию и выдает списки конкордансов для каждой из них. В то же время она подсчитывает общую частотность для каждого типа номинализации в каждом регистре, что позволяет исследовать каждый тип номинализации в контексте.

Специфические номинализации, часто встречающиеся в регистре, зависят от тем, затронутых в текстах корпуса. Так, в научной литературе шесть номинализаций встречаются существенно чаще других с частотами более 500 на млн слов: movement (почти 900 случаев встречаемости на млн слов), activity, information, development, relation и equation. Напротив, ни одна из номинализаций не встречается достаточно часто ни в художественной прозе, ни в устной речи. Например, movement встречается около 100 раз на 1 млн слов в художественной прозе и около 60 раз в устной речи, development встречается всего 10 раз на 1 млн слов в художественной прозе и практически не зафиксировано в устной речи.

Анализ конкордансов для этих шести номинализаций показывает, что в научной литературе номинализации описывают действия и процессы как абстрактные объекты, отделенные от человеческого участия. Эта модель видна на примере номинализации movement: The legs and

hips, or arms and shoulders, may be used to initiate movement in any direction. Движение в данном контексте – это процесс, представленный с помощью существительного, которое можно использовать как подлежащее или дополнение в частях сложного предложения.

В текстах научной литературы обсуждается обобщенное действие перемещения, а не перемещение какого-либо субъекта. Художественная проза и устная речь, с другой стороны, больше обращены к человеку, поэтому в этих регистрах чаще употребляются глаголы и прилагательные, чтобы описать поведение людей. Так, эти регистры часто имеют в качестве субъектов действия конкретных людей, и в них часто употребляется глагол move: Garth whistled breathily to himself and moved his hand crabwise along the table (fiction); It's how much they move it that counts (spoken).

Эту же модель можно увидеть в использовании таких номинализаций, как activity, development и information: The third Important aspect of information is speed; Sometimes algae can stop the development and growth of these plants; The experimental results can be described quantitatively by defining the size and activity of the shoot and root systems. В художественной прозе и в устной речи те же процессы и действия представлены с помощью глаголов или прилагательных, описывающих то, что делают определенные люди: I do hope you know that never in this country do we develop the sort of mob war that makes a protest against something however unjust develop into an organized riot (spoken); I've informed the Soviet government of that visit (spoken); "Aye, the big fellow is active again you'll be pleased to know." (fiction).

Эти обобщения приведены здесь для того, чтобы сказать, что существует ассоциативная связь между регистрами и распределением номинализаций. Научная литература намного чаще говорит о статической номинализации, в то время как художественная проза и устная речь описывают действия конкретных людей с помощью глаголов и прилагательных.

Распределение и функция суффиксов номинализаций

После исследования номинализаций как группы важно узнать о том, как распределяется каждый суффикс в отдельности, и, следовательно, о функциях разных типов номинализаций и роли, которую они играют в разных регистрах. Табл. 17, не показывая сами частоты, демонстрирует относительные пропорции номинализаций с каждым суффиксом в каждом регистре.

Таблица 17. Пропорции номинализаций с каждым суффиксом

Регистр / Суффикс	Научная литература	Художественная проза	Устная речь
-tion/-sion	68 %	51 %	56 %
-ment	15 %	21 %	24 %
-ness	2 %	13 %	5 %
-ity	15 %	15 %	15 %

1. Хотя суффикс -tion/-sion встречается в большинстве номинализаций во всех трех регистрах, его пропорция намного выше в научной литературе (68 %).

2. Суффикс -ment встречается в большем количестве в устной речи и художественной прозе, чем в научной литературе.

3. Суффикс -ness чаще встречается в художественной прозе, чем в каждом из двух других регистров.

Хотя нет абсолютных правил, отвечающих за выбор среди суффиксов номинализаций, тщательный анализ каждого типа выявляет определенные системные различия в значении, проливая свет на модели распределения, показанные в табл. 17. Например, суффиксы -tion/ -sion используются для преобразования глагола в существительное, обычно обозначающее обобщенный процесс или состояние (relate/relation и educate/education), поэтому в научной литературе отмечен самый высокий процент номинализаций с этими суффиксами.

Суффикс -ment также используется для преобразования глагола в существительное. Исчисляемые существительные, образованные с помощью суффикса -ment, часто обозначают процессы производства чего-либо или активности. Многие из этих номинализаций встречаются во всех трех регистрах, например, movement, government, achievement, agreement, argument. Тем не менее, многие номинализации с суффиксом -ment не являются исчисляемыми, обозначая ментальные состояния: amazement, agreement, astonishment, disappointment, embarrassment, excitement. Эти виды номинализаций являются редкими в научной литературе и устной речи, но в художественной прозе они довольно часто встречаются для описания ментального состояния персонажей: Patrick shrugged in embarrassment; The assembly cried out savagely and Ralph stood up in amazement.

Эти же номинализации иногда употребляются и в устной речи: I can quite see there’s cause for disappointment. Однако чаще ментальные

состояния в устной речи обозначаются глаголами и прилагательными: you'll be amazed; are you disappointed by not getting honors yourself?

Хотя в художественной прозе отмечено наименьшее количество номинализаций, нормированные подсчеты для номинализаций с суффиксом -ness являются самыми большими в этом регистре (1430 на 1 млн слов против 890 в научной литературе и 480 в устной речи). Суффикс -ness *обычно* преобразует прилагательные в существительные, обозначающие личные качества. В художественной прозе количество существительных, оканчивающихся на -ness больше, чем в других регистрах: awareness, bitterness, goodness, happiness, politeness, weakness. Эти слова важны для детального описания, которое характерно для художественной прозы: The bitterness in his heart was now mixed with a kind of childlike excitement; He could see Phyllis's face in profile, and it radiated energy and happiness.

В устной речи для описания личных качеств и чувств употребляются прилагательные, часто в роли определений к существительному, обозначающему говорящего или адресата: We feel frustrated and bitter and annoyed; I'm not too happy about this lauding of language...

Об употреблении номинализаций можно было бы сказать гораздо больше. Можно было бы рассмотреть количество разных слов с каждым суффиксом, чтобы определить, какой из них является наиболее продуктивным; можно было бы предпринять диахроническое исследование и проследить за развитием и употреблением номинализаций. Однако уже данный пример иллюстрирует силу основанного на корпусных данных подхода к морфологическим исследованиям. Очевидно, что деривационные суффиксы имеют ассоциативные связи с определенными регистрами, отражая первостепенные коммуникативные функции регистров.

3.3.4.2. Распределение грамматических категорий

Поскольку научная литература фокусирует внимание на абстрактных состояниях, процессах и объектах, а художественная проза включает более личные описания и действия, выполняемые конкретными людьми, должны быть различия в частотах существительных и глаголов во всех трех регистрах. Действительно, исследователи, характеризуя стили определенных авторов, иногда использовали сравнительные подсчеты существительных и глаголов. Такие подсчеты, по мнению авторов учебника Corpus Linguistics, могут также показать разницу по регистрам.

Частотность грамматических категорий

В морфологически аннотированном корпусе подсчет частоты существительных и глаголов являются относительно легким делом. Каждое существительное будет иметь тег, начинающийся с “n”, а каждый глагол будет иметь тег, начинающийся с “v”. Потому легко написать новую или использовать готовую программу для подсчета слов, маркированных одним из этих тегов. Более сложным является вопрос о том, что именно относить к глаголам и к существительным (в первую очередь, это проблема разметки):

1. Считать ли существительные, которые употребляются в роли определения последующего существительного? Например, должны ли словосочетания grasshopper ecology и animal groups подсчитываться как одно существительное или как два? С одной стороны, grasshopper и animal служат для определения следующих за ними существительных, и в этом отношении они подобны прилагательным в словосочетаниях типа general ecology или small groups. С другой стороны, такое словосочетание как grasshopper ecology содержит два отдельных референта – кузнечики (grasshoppers) и экология (ecology). В этом смысле они отличаются от фраз с прилагательным в качестве определения. Для подсчета количества глаголов и существительных представляется правильным *считать* существительные, которые определяют другие существительные.

2. Еще одна проблема касается местоимений. Если местоимения замещают существительные, в каком-то смысле они обозначают сущность или абстракцию. Однако они отличаются от существительных тем, что ничего не обозначают, если употребляются изолированно. Например, референт для слова he не может быть идентифицирован без специфического контекста, в отличие от слова grasshopper. Если нужно подсчитать слова, непосредственно относящиеся к предметам, местоимения подсчитывать *не следует*.

3. Подобные вопросы возникают и при подсчете глаголов. Например, нужно ли включать вспомогательные глаголы в общий подсчет глаголов? Следующие предложения из художественной прозы включают вспомогательный глагол в дополнение к смысловому глаголу: He had left home a little before eight; Joanne and her mother were talking. Эти вспомогательные глаголы не передают никакого лексического содержания. Вместо этого, они служат только для маркировки аспектуального значения (perfect или progressive) или требуются для построения негативной конструкции. Таким образом, целесообразно *исключить*

вспомогательные глаголы из общего подсчета относительной встречаемости глаголов и существительных.

Следовательно, принятие принципиального решения в каждом конкретном случае – это важная задача для составителя корпуса. Для пользователя важно также определить, как проводились подсчеты в других исследованиях, прежде чем сравнивать полученные результаты с результатами других исследований, потому что разные способы подсчетов дают разные данные о том, насколько предметным («nouny») является регистр. Табл. 18 показывает соотношение «существительное/глагол» при использовании трех разных способов подсчета существительных и глаголов.

Таблица 18. Соотношение «существительное/глагол» в трех регистрах

	Регистр / Категория	Научная литература	Художественная проза	Устная речь
А	Все существительные и глаголы	2,2	1,2	1,2
Б	Все существительные и глаголы, за исключением вспомогательных	2,9	1,5	1,6
В	Существительные, за исключением использованных в роли определения, и глаголы, за исключением вспомогательных	2,5	1,3	1,3

Строка А включает все существительные и глаголы, строка Б не включает вспомогательные глаголы, а строка В не включает существительные в роли определений для других существительных, равно как и вспомогательные глаголы. Местоимения исключены из всех подсчетов существительных.

Для подсчетов в каждой строке общее количество существительных было разделено на общее количество глаголов, чтобы показать, сколько существительных встречается на каждый глагол. Прежде чем остановиться на одном из способов подсчета, обратим внимание, что относительное количество одинаково для всех трех подсчетов. Художественная проза и устная речь имеют одинаковое соотношение «существительное-глагол», в то время как в научной литературе это соотношение почти в два раза больше. С любым методом подсчета полу-

ченные по всем регистрам модели одинаковы, хотя точные соотношения в регистрах варьируют.

Сравнение соотношения «существительное/глагол» по регистрам

Авторами учебника Corpus Linguistics самым подходящим способом подсчета существительных и глаголов признан подход Б (см. табл. 18), т.е. исключение местоимений и вспомогательных глаголов из подсчетов. Соотношение «существительное/глагол» в научной литературе намного превысило соотношение в других регистрах (2,9 : 1 против 1,5 : 1), что может быть проинтерпретировано так же, как и результат исследования номинализаций (см. п. 3.3.4.1). Так, следующий пример из текста, представляющего научную литературу, содержит 10 существительных и только один глагол:

Пример 1. In planning a livestock building or conversion, the psychological and health requirements of the livestock should undoubtedly be **given** absolute priority together with the basic needs of the stockman.

В этом примере виден акцент научной литературы на объектах, состояниях и процессах, которые обозначены существительными. Вместо того чтобы описывать, как человек планирует конструирование здания, здесь используются предметные описания обобщенных процессов: planning, conversion.

Напротив, типичные примеры из художественной прозы и устной речи содержат намного больше непосредственных действий. Следующий пример содержит 7 существительных и 5 глаголов, описывающих действия определенного человека:

Пример 2. He **emerged** and **locked** the door. He **unsnapped** the protective strap on his holster and **scanned** the parking lot. He **walked** quickly to the glass door of the bank.

Короткий пример неформальной беседы содержит еще больше глаголов, чем существительных – 14 глаголов и 4 существительных:

Пример 3.

A: Oh yeah, it's **called washing** your hair. Don't you **know** how to **wash** your hair?

B: Might **be**.

C: I **know**, I **know** how to **have** a bath.

B: **Go** away, I'm **cooking**. . . . **Excuse** me please, I'm **trying** to **cook**. I haven't **got** enough potatoes.

В примерах из художественной прозы и устной речи местоимения занимают место многих существительных, что уменьшает соотноше-

ние «существительное/глагол». Так, в примере неформальной беседы больше местоимений, чем существительных: I, me, you и it встречаются в 8 раз чаще. В примере из научной литературы местоимений нет совсем.

Анализ примеров показывает, что существуют важные и системные модели употребления, ассоциируемые с грамматическими закономерностями на всех уровнях. Понимание этих моделей является ключевым для полного понимания грамматики. Грамматическая вариативность присуща всем человеческим языкам, и эмпирические исследования языкового употребления раскрывают функциональную подоплеку этих структурных вариантов. Эти исследования важны как для дидактических целей, так и чисто научных. Интуиция носителя языка не всегда является надежной в предсказании того, какой вариант более предпочтителен, чем другой. Только основанное на корпусных данных исследование реальных текстов подходит для выявления этих моделей [40].

3.3.5. Исследования дискурса, основанные на корпусах

Закономерности употребления многих лексических и грамматических явлений можно полностью понять только путем анализа их функций в больших дискурсивных контекстах.

Под дискурсом понимают связный текст в совокупности с экстралингвистическими – прагматическими, социокультурными, психологическими и другими факторами; текст, взятый в событийном аспекте; речь, рассматриваемую как целенаправленное социальное действие, как компонент, участвующий во взаимодействии людей и связанный с механизмами сознания (когнитивными процессами) [23]. Элементы дискурса: излагаемые события, их участники, перформативная информация и «не-события», т. е. обстоятельства, сопровождающие события; фон, поясняющий события; оценка участников событий; информация, соотносящая дискурс с событиями [12]. В развитие анализа дискурса значительный вклад могут внести методы корпусной лингвистики, которые позволят тщательно описать характеристики определенных типов дискурса и то, до какой степени отдельный текст соответствует моделям дискурса в данном регистре. Основанные на корпусах исследования дискурса могут быть разделены на 4 сферы: 1) организация дискурса и структура текста; 2) дискурсивно-прагматические аспекты взаимодействия; 3) текстуальные и прагматические коллокации; 4) вариативность в текстах и в дискурсе. Из этих сфер наиболее перспективными являются, по мнению Т. Виртанен [61], две последние,

так как именно они выиграют от широкомасштабных исследований, возможных с помощью корпусов, поскольку вариативность предполагает устоявшиеся и новые правила, а коллокации предоставляют доступ к осязаемым и прагматическим аспектам лексики, грамматики или границ предложений.

Авторы учебника Corpus Linguistics полагают, что двумя основными способами применения корпусного подхода к исследованию дифференциальных признаков дискурса являются следующие:

1) для анализа характеристик дискурса целесообразно разработать и использовать интерактивные компьютерные программы (подобные программам проверки орфографии). В отличие от человека, они способны намного быстрее и надежнее выявлять определенные свойства дискурса, в то же время позволяя исследователю самостоятельно принимать решения в случаях, не поддающихся автоматическому анализу;

2) для отслеживания употреблений поверхностно-грамматических дифференциальных признаков во всем тексте может быть использован автоматический анализ. Эти типы анализа действительно задают развитие дискурсивных моделей по всем текстам, их можно использовать для сравнения текстов, для выявления специфических моделей, свойственных определенным регистрам, и для того, чтобы увидеть, как конкретный текст соотносится с общими регистровыми моделями.

Для конкретного рассмотрения вышеупомянутых способов применения корпусного подхода к исследованиям в области дискурса необходимо решить ряд вопросов, например, каким образом осуществляется разметка такого явления, как референция в текстах различных типов. В учебнике Corpus Linguistics проводится исследование употребления существительных и местоимений в четырех регистрах (неформальная беседа и публичная речь из корпуса Лондон-Лунд и новостной репортаж и научная литература из корпуса Ланкастер-Осло-Берген (LOB) с целью найти ответ на следующие вопросы: 1) какие факторы влияют на выбор между существительными и местоимениями в тексте? 2) какие существительные представляют «данную» (или «известную») информацию, а какие представляют «новую» информацию? 3) как известные и новые референты распределяются по тексту?

Грамматические связи между предложениями, при помощи которых осуществляется устный и письменный дискурс, могут быть разделены на три типа: референция, эллипсис и союзы. Референция в английском языке включает личные местоимения (he, she, it, we, they и т. д.), указательные местоимения (this, that, these, those), определенный

артикль the и выражение *such a*. Грамматический прием замены существительного местоимением называется прономинализацией (pronominalization). Это местоимение может относиться к существительному, которое было упомянуто в тексте раньше, позже, либо выходить за рамки текста, но входить в контекст дискурса [50].

Выделяют 3 типа референции, если статус информации определяется как «известная»: анафорическую, экзофорическую и выводимую. При анафорической референции местоимение заменяет собой ранее упомянутое в тексте существительное, например: The room is large. It is light and clean. Экзофорическая референция представляет собой ссылку на существительное, которое находится за пределами текста, но подразумевается как часть той ситуации, в которой происходит действие, и входит в контекст дискурса. Экзофорическая референция не всегда эксплицитна, т.е. для ее правильного понимания необходимо быть в курсе происходящих событий, например: That winter. It was awful. При выводимой референции информация, требуемая для интерпретации референтного средства, находится в самом тексте.

Именные конструкции являются основным грамматическим средством, отсылающим к людям, объектам и другим сущностям в тексте. Однако тексты, относящиеся к разным регистрам, часто сильно различаются в использовании этих «отсылочных выражений». В учебнике Corpus Linguistics рассматривается два примера из новостного репортажа и неформальной беседы, в которых именные конструкции выделены курсивом:

Пример 1. Новостной репортаж

Thortec International Inc. said it reached agreements with an investor group and Wells Fargo Bank under which it will receive loans and an equity infusion in return for stock that will reduce the number of shares in public hands by as much as 85 percent. The engineering and consulting firm, which has been plagued by losses for five years, said the restructuring is required to relieve its debt burden and "acute shortage of cash."

Пример 2. Неформальная беседа

A: Right, I'm ready. Have you locked the back door? [pause] I thought we were walking.

B: Well do you want to walk or do you want to go in the car?

A: Well I have to go to the paper shop.

B: Well I'll drop you at the paper shop while I go round.

A: Oh that's a good idea.

Одно хорошо заметное различие между этими примерами текстов касается формы именных конструкций. В примере из новостного ре-

портажа в основном употребляются полные именные конструкции (Thortec International Inc., agreements, an investor group и др.), тогда как в примере из неформальной беседы более часто применяются местоимения (I, you, we, that). Кроме того, очевидно, что в этих примерах употребляются разные типы референции. В частности, в примере из неформальной беседы присутствует большой процент экзофорической референции с местоимениями I и you, напрямую связанными с говорящим и адресатом, а не с каким-либо объектом, ранее встретившимся в тексте. В примере из новостного репортажа такого типа референции нет. К тому же, из-за большей опоры на экзофорическую референцию, большее количество референтов в примере из неформальной беседы уже знакомо обоим участникам даже при первом упоминании о них, например: I, you, the back door, the paper shop, в то время как большее количество референтов в примере из новостного репортажа изначально незнакомы, например: agreements, an investor group.

Базирующийся на корпусных данных анализ может быть применен для исследования характеристик референциальных выражений и для определения степени различия их использования в разных регистрах.

Характеристики референциальных выражений

Существует много характеристик референциальных выражений, которые можно исследовать для того, чтобы лучше понять их употребление в разных текстах и регистрах. В учебнике Corpus Linguistics анализируется в качестве примера 4 параметра:

1) статус информации: известная, новая;

2) тип референции для известной информации: анафорическая, экзофорическая или выводимая;

3) форма выражения для анафорической референции: местоимение, синоним или повтор;

4) расстояние между анафорическим выражением и антецедентом для анафорической референции.

Каждая из именных конструкций в тексте может быть классифицирована в соответствии с типом представленной в ней информации – известной или новой. Так, в примере из новостного репортажа (пример 1) многие именные конструкции представляют новую информацию, указывая на человека или объект, ранее не упомянутый в тексте. Именные конструкции такого типа включают следующие: Thortec International Inc., an investor group, Wells Fargo Bank, loans, an equity infusion, stock. Другие референциальные выражения представляют известную инфор-

мацию, вводя сущность, которая уже была упомянута. Так, в первом предложении местоимение it употреблено дважды, чтобы обозначить известный референт, а именно: компанию Thortec International Inc.

Выражения, вводящие известную информацию, представляют три типа референциальных отношений. Многие из таких выражений являются анафорическими, т. е. относятся к человеку или объекту, уже упомянутому в тексте – антецеденту. Так, антецедентом для местоимения it в первом предложении является Thortec International Inc. Однако другие референты представляют известную информацию в силу того, что они относятся к человеку или объекту во внешнем контексте. В примере из неформальной беседы (пример 2) местоимения I и you прямо указывают на говорящего и адресата. The back door, the car и the paper shop относятся к физическим объектам, присутствующим в расширенной физической ситуации, которая понятна обоим участникам разговора. Такие референты называются экзофорическими. Они являются известными, поскольку их идентификация возможна благодаря *физической ситуации*. Напротив, анафорические референты известны потому, что их идентификация возможна благодаря предшествующей *текстовой референции*.

Существуют способы выражения известной информации, которые классифицировать еще сложнее. В частности, в примере из новостного репортажа (пример 1) существование restructuring, к которому обращаются во втором предложении, является «выводимым» из событий, описанных в первом предложении, но это существительное не относится анафорически ни к одной из предшествующих именных конструкций и не относится к внешнему контексту. Подобным образом, существование debt burden может быть выведено из того факта, что компания была "plagued by losses," но это также не является анафорическим отношением. Следовательно, категория «выводимый» также важна для классификации референтов.

Третий параметр касается различных форм представления анафорических референтов. Они часто выражаются местоимениями, однако могут быть выражены и синонимическими выражениями, например, the engineering and consulting firm во втором предложении относится к Thortec International. Кроме того, анафорические референты могут быть прямым повтором первоначального выражения. Четвертый параметр касается расстояния между референциальным выражением и его антецедентом. Так, в примере из новостного репортажа местоимение it оказывается относительно близко к антецеденту Thortec International Inc. Более полное синонимическое выражение The engineering and con-

sulting firm располагается на большем расстоянии от первоначального упоминания этой компании. Все четыре параметра вместе могут раскрыть многие модели использования референции в разных регистрах. Анализ даже нескольких тысяч слов текста может быть очень затратным по времени, поэтому для изучения характеристик референциальных выражений целесообразно использовать корпусный подход.

Техника интерактивного анализа: кодирование характеристик референциальных выражений

Чтобы проиллюстрировать результаты работы интерактивной программы по анализу текста с целью выявления референтных типов, авторами учебника Corpus Linguistics были обработаны первые 200 слов из 40 текстов, взятых из корпусов Лондон-Лунд и Ланкастер-Осло-Берген. Тексты были представлены четырьмя жанрами: неформальная беседа (5 текстов), публичная речь (9 текстов), новостной репортаж (10 текстов) и научная литература (16 текстов). Была разработана программа, направленная на выявление и анализ шести характеристик для каждой именной конструкции:

1) регистр, который предварительно указывается в начале каждого текста и не вовлекается в последующий анализ;

2) форма именной конструкции (местоимение или существительное), определяющаяся на основе процедуры аннотирования;

3) статус информации (новая или известная), причем местоимения автоматически рассматриваются как известная информация, а для каждого существительного осуществляется проверка, встречается ли оно в предшествующем фрагменте текста. Если да, то программа приписывает статус «известная», если нет, то предварительно отмечает информацию как новую, предлагая эксперту самому решать, правильно ли это;

4) тип референции (анафорическая, экзофорическая, выводимая), если статус информации определяется как «известная», причем местоимения I и you автоматически соотносятся с экзофорическим типом референции, а местоимения 3-го лица и существительные, рассматривающиеся как известная информация, размечаются программой как анафорические, но проверяются впоследствии в интерактивном режиме на экзофорическую и выводимую референцию;

5) тип выражения (синоним или повтор существительного), если представлен анафорический тип референции, выраженный существительным;

6) расстояние между антецедентом и референциальным выраже-

нием, вычисляемое как количество находящихся между ними именных групп.

Интерактивная программа анализа текста позволяет ускорить работу исследователя и обеспечивает более высокую точность данных. Сначала проводилась морфологическая разметка всех текстов, затем интерактивная программа обрабатывала каждый размеченный текст, останавливаясь на каждом местоимении и существительном, позволяя пользователю выбрать правильные коды для именных конструкций. Если первичный анализ информационных характеристик, автоматически проведенный программой, является правильным, пользователь просто принимает код, а если нет, то программа предоставляет список других вероятных вариантов анализа, из которых можно выбирать путем простого указания номера, соответствующего правильному варианту. На рис. 6 приводится пример работы программы, показывающий, как коды могут быть приняты или отредактированы. Референциальное выражение (them), которое подлежит кодированию, представлено в контексте и обозначено стрелкой.

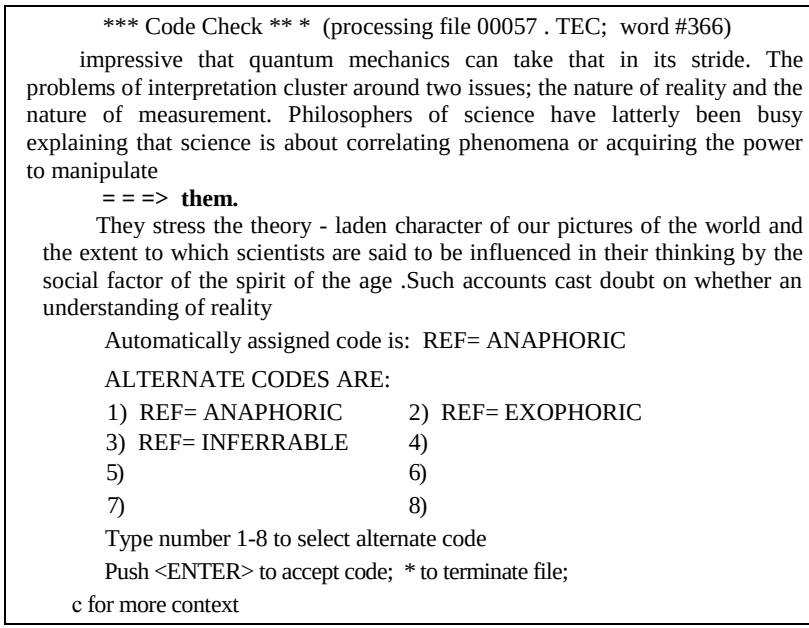


Рис. 6. Пример работы интерактивной программы кодирования референциальных выражений.

Под примером текста приведены автоматически присвоенный код ("ANAPHORIC") и альтернативные варианты кодов. Когда все именные конструкции проанализированы, коды записываются в тексте такими строками как: <<<Ref = anaphoric и <<<Status = given

Затем используется другая компьютерная программа для анализа кодированного текста и создания файла, перечисляющего информационные характеристики каждой именной конструкции. В конце проводится статистический анализ, показывающий взаимодействие этих характеристик.

Подобные результаты, полученные при помощи интерактивных компьютерных программ и автоматических методов обработки текста, имеют большое значение для анализа дискурса. Особенно целесообразно использование корпусного подхода для выявления характеристик дискурса, присущих тому или иному регистру [40, 33].

Распределение обращений в неформальной беседе

Исследование дискурса, предпринятое Дж. Личем, было посвящено распределению и функционированию обращений в беседе на американском и британском вариантах английского языка [48]. Объектом основанного на корпусе исследования была грамматика обращений, понимаемых как субстантивные свободно присоединяемые элементы, не являющиеся членами предложения, относящиеся к адресату высказывания.

Изучая данные собранного добровольцами корпуса, автор выделяет несколько семантических подкатегорий: ласковое обращение: Honey, can I use that ashtray, please; обращение к родственникам: Thanks, mom, ok, talk to you later; «фамильяризующее» обращение (familiariser): Got a ticket, mate?

Автор подсчитывает все случаи встречаемости по всем подкатегориям и выявляет следующие различия между британским и американским вариантами английского языка: в американском варианте обращения используются на 25 % чаще, чем в британском, термины родства чаще встречаются в британском варианте, «фамильяризующие» обращения чаще употребляются в американском варианте.

В работе исследуются обращения с точки зрения их места в предложении (табл. 19), а также различные отношения между типом функции (привлечение внимания, указание на адресата, усиление социального взаимодействия) и позицией обращения в предложении.

Таблица 19. Место обращения в предложении

Место в предложении	Кол-во, %	Пример
Конец предложения	68%	Come on, Sam.
Начало предложения	11,5%	Doug, do you want some more ice-cream?
Отдельное расположение	11,25%	Mom!
Середина предложения	9, 25%	What have we lost at home, Paulie, this season?

Автору удалось получить новые результаты:

1. Люди используют обращение sir в разговоре с официантами, что, возможно, объясняется сдвигом общества в направлении демократизации.

2. Обращения чаще всего встречаются в конце предложения, что может быть объяснено большей важностью их социальной функции по сравнению с другими, включая функцию привлечения внимания.

Исследование дискурса на материале звукового корпуса

На материале русского языка проводилось исследование энантиосемии – совмещения в слове противоположных значений, «внутренней антонимии» [24]. Рассмотрение этого явления на материале звукового корпуса русского языка повседневного общения *Один речевой день* (ОРД) (см. п. 3.2.2.2.5) привело автора к мысли о том, что есть прежде не выявлявшийся лингвистами тип энантиосемии, часто используемый в разговорной речи, а именно: риторическая энантиосемия.

Наблюдения над лексикой определенных фрагментов корпуса ОРД показали, что информант И19 (женщина 30–35 лет, в общей системе обозначения информантов в базе «Один речевой день» именуемая И19), не говорит ничего обидного, желает собеседнику удачи, даже хвалит (*молодец, замечательно, хорошо, отлично*). Однако при этом имя ребенка и другие маркеры доверительности (типа *солнышко, заинька, зайка, котик, умница, дорогой, милый, миленький и т. д.*) произносятся без характерных для разговора с детьми и свойственных доброжелательному адресанту особенностей: нет ни продления долготы звука сверх обычного для ударных, ни варьирования частоты основного тона, ни повышения регистра, ощущается отсутствие эмоциональной составляющей.

В следующих далее фрагментах текста адресант использует обращения, традиционные для разговоров с близкими людьми (*солнышко,*

(мое) солнце, (моя) радость), а также само имя ребенка с уменьшительно-ласкательными суффиксами *оньк, очк*:

1) *Але-о!/ Привет / что все?/ закончилось / у вас закончилось /что случилось?/ почему? / ты где? т. е. она сейчас уже ушла? а спроси у Иры /да-а/ я же тебе сказала /подойти к Ирине и спросить/ попросить помощи / Да / дойди / дойди* ***солнышко****/ давай/ удачи/ ну тебе/ у тебя все хорошо/ ну замечательно/ ну /угу /хорошо /****солнышко*** */ давай /не теряй времени /подойти к Ирине /попроси помощи найти Нину Филипповну/и отзвонись мне/ пока //*

2) *Але /да* ***солнышко*** */ну так /ну замечательно/ отлично /поздравляю тебя /* ***молодец*** */все не так страшно /готовься /что делать/ Держись /держись мое* ***солнце*** */ну давай / держись / пока //*

3) *Ты выпила* ***водичку****? Отлично! <…> Что случилось? Что именно ты забыла, моя* ***радость****? Можно поподробнее? Что ты забыла? Что сегодня у вас был английский/ ты забыла. <…>Я не поняла, что да? Зачет?<…>проверка по словам<…> Какие слова? Что/ и сейчас забыла? И сейчас забыла / какие слова? <…> Из какой лексики/* ***Лизонька*****?**

4) ***Лизонька*** */ это Марина Викторовна ваша / такое дурацкое слово употребляет «лексика» /оно дурацкое /* ***Лизочка*** */ Лексика это всего-навсего слова <…>Лексику вы учите / бред какой /*

5) ***Моя девочка/*** *<…>ну что делать-то с этой двойкой /****моя хорошая***

Была выдвинута гипотеза о том, что особенности интонирования маркеров доверительности создают конфликт горизонта ожиданий слушающего и интенций говорящего, а диссонанс между семантикой слова и нейтральной интонацией (при ожидаемой экспрессивной) становится основой для аномального эмоционального фона при общении. Для проверки этой гипотезы был проведен эксперимент, участникам которого (группа из 30 студентов и школьников) было предложено *прочитать* расшифровку и ответить на такие вопросы: «Часто ли родители так разговаривают с детьми? Типично ли содержание разговора? Что можно сказать об этой женщине?» Информанты восприняли текст как банальный, многие предположили, что ребенок должен сдавать какой-то экзамен или зачет, которого боится, что в связи с этим мама очень переживает, волнуется, старается поддержать дочку.

После *прослушивания* записи предлагалось ответить на вопрос: «Что нового вы узнали о Лизе?» Информанты реагировали крайне эмоционально, причем не отвечали на поставленный вопрос, а выражали мнение по поводу высказываний И19: «Почему она таким прокурор-

ским тоном разговаривает?» «А Вы можете на нее повлиять, чтобы она так больше не говорила?» «Вот пойдет в школу у Вас сын, Вы, может, тоже еще так заговорите!» «Эта женщина, наверное, просто устала, а дочка у нее еще неизвестно что такое.» Реплики, при всем разнообразии оценок поведения Лизы и ее матери, отражают одно: в прослушанных фрагментах участники эксперимента ощутили нечто неприятное, раздражающее, чего невозможно уловить при письменной передаче текстов.

Значения слов с риторической энантиосемией, участвующих в коммуникативных актах, не претерпевают никаких трансформаций: не происходит ни мелиорации, ни пейоративизации значений, не актуализируются непрямые значения. Вместе с тем, здесь нет и интонационного отрицания называемых признаков или фактов. Даже лишенная интонации угрозы или иронии, фраза, включающая риторическую энантиосемию, создает напряженный эмоциональный фон. Не случайно при прослушивании звукового файла продолжительностью 22 мин. 36 сек. (продолжительность разговора с матерью) девочка пытается заплакать семь раз. На записи слова с положительной коннотацией, требующие эмоционально окрашенной интонации, произносятся нейтральным тоном. Многочисленные повторы ласковых слов (примеры 5, 6, 7), произносимых таким образом, способствуют нарастанию напряжения, а затем разрешаются каскадом вопросов или жалобами.

Особую интонацию отстраненности, отчужденности, которая характеризует примеры (3–7), автор относит к маркерам чуждости. При этом в записях не наблюдается таких проявлений агрессивности речевого поведения, как сверхполный тип произношения, понижение тона, повышение голоса с целью оказать давление на собеседника [21]. В отличие от нормального (не агрессивного) речевого поведения, при котором собеседники чувствуют себя равноправными, в условиях речевой агрессии исключается равноправие участников диалога, один из них становится агрессором, другой – жертвой. Видимо, в репликах И19 и проявляется агрессия ради агрессии, с помощью которой за счет близких людей снижается эмоциональное напряжение говорящего [24].

Данный пример исследования показывает, что на материале звукового корпуса могут быть сделаны серьезные теоретические выводы. Автору удалось выявить новый тип энантиосемии – риторическую энантиосемию, которая заключается в совмещении в слове (словосочетании, предложении) контактоустанавливающей и деструктивно-

агрессивной коммуникативных установок, при котором семантика рассматриваемой единицы (включая оценочную составляющую) не меняется.

Вопросы для самоконтроля

1) Дайте определение следующим понятиям:
   treebank, *конкорданс*, *нормированная частота*, *регистр*, *коллокат*, *коллокация*
2) Что такое язык регулярных выражений?
3) Назовите и охарактеризуйте наиболее известные национальные корпусы.
4) Какие лингвистические особенности регистра художественной прозы можно выявить с помощью корпусной методологии?
5) Почему нелингвистические корпусы (базы данных поисковых систем) можно рассматривать как корпусы текстов?

ЗАКЛЮЧЕНИЕ

Корпусная лингвистика представляет собой новое направление в лингвистической науке, позволяющее проводить исследование единиц любого языкового уровня в реальном их употреблении, т. е. с учетом того, в какой ситуации то или иное высказывание было произведено. Большие национальные корпусы и корпусы, созданные для специальных целей, позволяют исследователям осуществлять автоматический поиск и систематизацию эмпирического материала, быстро обрабатывать большие массивы языковых данных.

Это одна из стремительно развивающихся областей, и если считать, что корпусная лингвистика – это, в первую очередь, методология проведения лингвистических исследований, необходимо подчеркнуть, что прогресс в сфере компьютерных технологий влечет за собой прогресс в создании и совершенствовании программ автоматической обработки текста и, как результат, порождает новые парадигмы лингвистических исследований.

Авторы отдают себе отчет в том, что часть сведений, приведенных в учебнике, со временем устареет, что на смену тем или иным конкретным программам придут более совершенные и многофункциональные, появятся новые корпусы, изменятся их адреса в сети Интернет. Свою задачу авторы видели в том, чтобы дать описание нового направления в лингвистике как такового и состояние корпусной лингвистики в начале второго десятилетия XXI в.

ПРИМЕРНАЯ ТЕМАТИКА ДОКЛАДОВ, РЕФЕРАТОВ, КУРСОВЫХ РАБОТ

1. Способы использования корпусов в лингвистических исследованиях.
2. Исследование способов использования корпусов в лексикографии.
3. Изучение средств обработки корпусных данных, представленных на языке XML.
4. Создание электронной хрестоматии по корпусной лингвистике.
5. Исследование механизмов взаимодействия корпуса текстов и электронной картотеки (корпусы цитат).
6. Создание веб-сайта по корпусной лингвистике.
7. Графематический анализ текстов.
8. Унификация текстов внутри корпуса.
9. Автоматическая морфологическая разметка текстов XIX в.
10. Исследование набора метаданных для корпуса XIX в.
11. Создание параллельного корпуса.
12. Методы снятия морфологической неоднозначности.
13. Анализ функций сегментных внеалфавитных графем («межморфемный» дефис, «межслоговой» дефис, «межсловный» дефис, апостроф).
14. Проблема строчных и прописных букв в корпусах текстов (имена собственные и нарицательные, сплошная и начальная капитализация).
15. Проблема омографии – акцентно-ориентированный морфологический анализ.
16. Разработка модуля преобразования каллиграфем (жирность, курсивность, подчёркивание) в тэги языка XML.
17. Анализ функций точки (и других знаков препинания) с точки зрения структурной разметки текста.
18. Методы выделения структурных элементов текста (часть, глава, параграф, абзац).
19. Составные лексемы.
20. Методы снятия морфологической неоднозначности.

21. Методы выделения структурных элементов текста (часть, глава, параграф, абзац).
22. Составные лексемы.
23. Проект TEI (обзор).
24. Стандарты EAGLES (обзор).
25. Форматы CDIF и XCES (обзор).
26. Анализ и описание различных корпусов.
27. Анализ и описание корпусного менеджера Xaira.
28. Анализ и описание корпусного менеджера Bonito.
29. Анализ и описание корпусного менеджера CQP.
30. Анализ и описание интерфейса WebCorp.
31. Сравнительный анализ возможностей корпусов и поисковых систем Интернета.
32. Использование корпусов в социологии.
33. Использование корпусов в этнолингвистике.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Основная

Баранов А.Н. Введение в прикладную лингвистику. – М., 2007.

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

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

Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: Учеб. пособие. – М.: Издательский центр «Академия», 2004.

Дополнительная

Азарова И.В., Алексеева К.Л., Захарова Л.А. Разметка текстовых фрагментов в корпусе агиографических текстов СКАТ // Труды международной конференции «Корпусная лингвистика – 2006». – СПб.: Изд-во С.-Петерб. ун-та, Изд-во РХГА, 2006. – С. 16–24.

Беляева Л.Н. Лексикографический потенциал параллельного корпуса текстов // Труды международной конференции «Корпусная лингвистика – 2004». – СПб., 2004. – С. 55–64.

Гарабик Р., Захаров В.П. Параллельный русско-словацкий корпус // Труды международной конференции «Корпусная лингвистика – 2006». – СПб., 2006. – С. 81–87.

Герд А.С. РНК и академическая лексикография // Труды международной конференции «Корпусная лингвистика – 2006». – СПб.: Изд-во С.-Петерб. ун-та; Изд-во РХГА, 2006. – С. 88–91.

Герд А.С. Несколько слов о Специальном корпусе текстов (СКТ) // Труды международной конференции «Корпусная лингвистика – 2006». – СПб.: Изд-во С.-Петерб. ун-та; Изд-во РХГА, 2006. – С. 92–93.

Гришина Е.А., Савчук С.О. Национальный корпус русского языка как инструмент для изучения вариативности грамматических норм // Труды международной конференции «Корпусная лингвистика – 2008» 6–10 октября 2008 г. – СПб., 2008. – С. 161–169.

Захаров В.П. Веб-пространство как языковой корпус // Компьютерная лингвистика и интеллектуальные технологии: Труды междуна-

родной конференции «Диалог-2005» (Звенигород, 1–6 июня 2005 г.) – М., 2005. – С. 166–171.

Камшилова О.Н. Учебный корпус текстов: потенциал, состав, структура. – СПб.: ООО «Книжный Дом», 2012. – 58 с.

Митрофанова О.А., Грачкова М.А., Шиморина А.С. Автоматическая классификация лексики в параллельных текстах (на материале текстов из Русско-словацкого корпуса параллельных текстов PARUS) // Материалы V Международной научно-практической конференции «Прикладная лингвистика в науке и образовании: лингвистические технологии и инновационная образовательная среда». – СПб.: «ЛЕМА», 2010. – С. 231–235.

Перцов Н.В. О роли корпусов в лингвистических исследованиях // Труды международной конференции «Корпусная лингвистика – 2006». – СПб.: Изд-во С.-Петерб. ун-та; Изд-во РХГА, 2006. – С. 318–331.

Потапова Р.К. Новые информационные технологии и лингвистика : учебное пособие. – М.: Едиториал УРСС, 2005.

Сичинава Д.В. Национальный корпус русского языка: очерк предыстории. 2005. – http://ruscorpora.ru/sbornik2005/03sitch.pdf

Труды международной конференции «Корпусная лингвистика – 2011» 27–29 июня 2011 г., Санкт-Петербург. – СПб.: СПбГУ. Филологический факультет, 2011. – 348 с.

Шерстинова Т.Ю. «Один речевой день» на временной шкале: о перспективах исследования динамических процессов на материале звукового корпуса // Филология. Востоковедение. Журналистика. Серия 9. – СПб., 2009.

Biber D., Conrad S., Reppen R. Corpus Linguistics. Investigating language structure and use. Cambridge University Press, 1998.

Lüdeling A., Kytö M., eds. Corpus Linguistics. An International Handbook. Volumes 1, 2. – Berlin & New York: Walter de Gruyter, 2008. – http://alknyelvport.nytud.hu/muhelyek/elte.../HSK-Corpus-Linguistics.../ file

Meyer Ch. F. English Corpus Linguistics: An Introduction. Cambridge: Cambridge University Press, 2002.

СПИСОК ИСТОЧНИКОВ ЦИТАТ

1. *Азарова И.В., Алексеева К.Л., Захарова Л.А.* Разметка текстовых фрагментов в корпусе агиографических текстов СКАТ // Труды международной конференции «Корпусная лингвистика – 2006». – СПб: изд-во С.-Петерб. ун-та, Изд-во РХГА, 2006. – С. 16–24.

2. *Баранов А.Н.* Введение в прикладную лингвистику. – М., 2007.

3. *Баранов А.Н., Плунгян В.А., Рахилина Е.В.* Путеводитель по дискурсивным словам русского языка. – М., 1993.

4. *Беляева Л.Н.* Лексикографический потенциал параллельного корпуса текстов // Труды международной конференции «Корпусная лингвистика – 2004». – СПб., 2004. – С. 55-64.

5. *Богданова С.Ю.* Исследование слова и предложения компьютерными методами // Слово в предложении: кол. монография / Под ред. Л.М. Ковалевой (отв. ред.), С.Ю. Богдановой, Т.И. Семеновой. – Иркутск: ИГЛУ, 2010. – С. 194-213.

6. *Борисова Е.Г.* Слово в тексте. Словарь коллокаций (устойчивых словосочетаний) русского языка с англо-русским словарем ключевых слов. – Москва, 1995.

7. *Гарабик Р., Захаров В.П.* Параллельный русско-словацкий корпус // Труды международной конференции «Корпусная лингвистика – 2006». – СПб., 2006. – С. 81-87.

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

9. *Герд А.С.* РНК и академическая лексикография // Труды международной конференции «Корпусная лингвистика – 2006». – СПб.: Изд-во С.-Петерб. ун-та; Изд-во РХГА, 2006. – С. 88-91.

10. *Герд А.С.* Несколько слов о Специальном корпусе текстов (СКТ) // Труды международной конференции «Корпусная лингвистика – 2006». – СПб.: Изд-во С.-Петерб. ун-та; Изд-во РХГА, 2006. – С. 92–93.

11. *Гришина Е.А.* Мультимедийный русский корпус: современное состояние и перспективы развития // Труды международной конференции «Корпусная лингвистика – 2011» 27–29 июня 2011 г. – СПб. СПбГУ, Филологический факультет, 2011. – С. 138–144.

12. *Демьянков В.З.* Англо-русские термины по прикладной лингвистике и автоматической переработке текста. Вып. 2. Методы анализа текста // Всесоюзн. центр переводов. Тетради новых терминов, 39. – М., 1982.

13. *Засорина Л.Н.* (ред.). Частотный словарь русского языка. – М., 1977.

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

15. *Захаров В.П.* Веб-пространство как языковой корпус // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2005» (Звенигород, 1–6 июня 2005 г.) – М., 2005. – С. 166–171.

16. *Зализняк Анна А., Левонтина И.Б., Шмелев А.Д.* Ключевые идеи русской языковой картины мира: Сб. статей. – М.: Языки славянской культуры, 2005.

17. *Зубов А.В., Зубова И.И.* Информационные технологии в лингвистике: Учеб. пособие. – М.: Издательский центр «Академия», 2004. *Камшилова О. Н.* Учебный корпус текстов: потенциал, состав, структура. – СПб.: ООО «Книжный Дом», 2012.

18. *Карлсон Ф.* Ранняя генеративная лингвистика и эмпирическая методология // Когнитивные категории в синтаксисе: Кол. монография. – Иркутск: ИГЛУ, 2009. – С. 215–247.

19. *Карпова О.М.* Английская лексикография : Учебн. пособие для студентов высших учебных заведений, обучающихся по специальностям направления «Лингвистика и межкультурная коммуникация». – Москва: Академия, 2010. – 174 с.

20. *Крейдлин Г.Е.* Голос и тон в языке и речи // Язык о языке / Отв. ред. Н.Д. Арутюнова. М.: Языки русской культуры, 2000. – С. 453–501.

21. *Крейдлин Г.Е.* Невербальная семиотика. – М.: Новое литературное обозрение, 2002. – 581 с.

22. *ЛЭС* – Лингвистический энциклопедический словарь / Под ред. В.Н. Ярцевой. – М.: Сов. энциклопедия, 1990.

23. *Маркасова Е.В.* Риторическая энантиосемия в корпусе русского языка повседневного общения «Один речевой день» // Компьютерная лингвистика и интеллектуальные технологии. Вып. 7 (14). По материалам ежегодной международной конференции «Диалог (2008) / Гл. ред. А. Е. Кибрик. – М., 2008. – С. 352–355.

24. Машинный фонд русского языка: идеи и суждения, – М.: Наука, 1989.

25. *Митрофанова О.А., Грачкова М.А., Шиморина А.С.* Автоматическая классификация лексики в параллельных текстах (на материале текстов из Русско-словацкого корпуса параллельных текстов PARUS) // Материалы V Международной научно-практической конференции «Прикладная лингвистика в науке и образовании: лингвистические технологии и инновационная образовательная среда». – СПб.: «ЛЕМА», 2010. – С. 231–235.

26. *Мордовин А.Ю.* К вопросу о жанровой полноценности современных неспециализированных корпусов текстов // Вестник ИГЛУ, 2009. – № 2. – С. 48–52.

27. *Национальный корпус русского языка*. – http://ruscorpora.ru

28. *Николаева Ю.В.* Кинетические признаки структуры устного нарратива (корпусное исследование) // Проблемы компьютерной лингвистики: Сборник научных трудов / Под ред. А.А. Кретова. – Вып. 4. – Воронеж, 2010. – С. 193–200.

29. *Перцов Н.В*. О роли корпусов в лингвистических исследованиях // Труды международной конференции «Корпусная лингвистика – 2006». – СПб.: Изд-во С.-Петерб. ун-та; Изд-во РХГА, 2006. – С. 318-331.

30. *Потапова Р.К.* Новые информационные технологии и лингвистика : Учебн. пособие. – М.: Едиториал УРСС, 2005.

31. *Рыков В.В.* Корпус текстов как реализация объектно-ориентированной парадигмы // Труды Международного семинара Диалог-2002. – М.: Наука, 2002.

32. *Сичинава Д.В.* Национальный корпус русского языка: Очерк предыстории. 2005. – URL: http://ruscorpora.ru/sbornik2005/03sitch.pdf

33. *Толпегин П.В.* Автоматическое разрешение кореференции местоимений третьего лица русскоязычных текстов: Автореф. дис. … канд. техн. наук. – М., 2008.

34. *Труб В.М.* О возможном подходе к семантическому описанию частей тела // Московский лингвистический журнал. – 2006. – Т. 8. – № 1. – С. 67–73.

35. *Асиновский А. С., Богданова Н. В., Русакова М. В., Степанова С. Б., Шерстинова Т. Ю.* Звуковой корпус русского языка повседневного общения «Один речевой день»: концепция и состояние формирования // Компьютерная лингвистика и интеллектуальные технологии. Вып. 7 (14). По материалам ежегодной международной конференции «Диалог» (2008). М., 2008. С. 488-494. (http://www.dialog-21.ru/digests/ dialog 2008/materials/ html/76.htm)

36. *Хохлова М.В.* Экспериментальная проверка методов выделения коллокаций // Slavica Helsingiensia 34. Инструментарий русистики: Корпусные подходы. – Хельсинки, 2008. – С. 343–357.

37. *Шаров С.А.* Частотный словарь русского языка. 2002. http://www.artint.ru/projects/frqlist.asp

38. *Шерстинова Т.Ю.* «Один речевой день» на временной шкале: о перспективах исследования динамических процессов на материале звукового корпуса // Филология. Востоковедение. Журналистика. Сер. 9. – СПб., 2009. С. 227–235.

39. Atkins S., Clear J., Ostler N. Corpus Design Criteria. Literary and Linguistic Computing. – 1992. – Vol. 7, No. 1. – P. 1–16.

40. Biber D., Conrad S., Reppen R. Corpus Linguistics. Investigating language structure and use. Cambridge University Press, 1998.

41. *Český národní korpus* – úvod a příručka uživatele FF UK / Kocek J., Kopřivová M., Kučera K. – Praha: ÚČNK 2000.

42. Finegan E. LANGUAGE: its structure and use. – N.Y.: Harcourt Brace College Publishers, 2004.

43. Johannessen J.B. Corpus linguistics or corpora in linguistics? // NODALIDA 2005 (The 15th Nordic Conference of Computational Linguistics. Joensuu, Finland, May 20–21, 2005).

44. Johansson S. Some aspects of the development of corpus linguistics in the 1970s and 1980s // Lüdeling A., Kytö M., eds. Corpus Linguistics. An International Handbook. Vol. 1. – Berlin; N.Y.: Walter de Gruyter, 2008. – P. 33–53.

45. Kilgarriff A. Web as corpus // Proc. of Corpus Linguistics 2001 conference (Lancaster University). – Lancaster, 2001. – P. 342–344.

46. *Kytö M., Rissan*en M. A language in transition: The Helsinki Corpus of English texts, ICAME Journal, 1992. Vol. 16: P. 7–27.

47. Lakoff G. Pronominalization, Negation, and the Analysis of Adverbs // Readings in English transformational grammar, Ginn & Co, Waltham, MA, 1970 / Eds. R. Jacobs, P. Rosenbaum. – P. 145–165.

48. Leech G. The Distribution and Function of Vocatives in American and British English Conversation // In: Out of Corpora. Studies in Honour of Stig Johansson // Eds. H. Hasselgård and S. Oksefjell, 1999. – P. 107–120.

49. *Lüdeling A., Kytö M.,* eds. Corpus Linguistics. An International Handbook. Vol. 1, 2. – Berlin; N. Y.: Walter de Gruyter, 2008. – http://alknyelvport.nytud.hu/muhelyek/elte.../HSK-Corpus-Linguistics.../file

50. McCarthy M.J. Discourse Analysis for Language Teachers. – Cambridge: Cambridge UP, 1991.

51. McEnery T., Wilson A. Corpus Linguistics. – Edinburgh: Edinburgh University Press, 2001.

52. McWhinney B. The CHILDES Project: Tools for Analyzing Talk. – Mahwah, NJ: Lawrence Erlbaum Associates, Inc. 3rd ed., 2000. – Vol. 1.

53. Meyer Ch. F. English Corpus Linguistics: An Introduction. Cambridge: Cambridge University Press, 2002. – Xvi + 168.

54. Mitrofanova O., Zacharov V. Automatic Analysis of Terminology in the Russian Corpus on Corpus Linguistics // Slovko-2009: NLP, Corpus Linguistics, Corpus Based Grammar Research: Proceedings of Fifth International Conference (Smolenice, Slovakia, 25–27 November 2009) / Eds J. Levicka, R. Garabik,. – Brno: Tribun, 2009. – P. 249-255.

55. Postal P.M. Cross-Over Phenomena. A Study in the Grammar of Coreference / Ed. W.J. Plath // Specification and Utilization of a Transformational Grammar. Scientific Report No. 3. P. 1–239. Yorktown Heights, N.Y: IBM Corporation, 1968.

56. Postal P.M. On the Surface Verb ‘remind’ // Linguistic Inquiry, 1970. – Vol. 1. P. 37–120.

57. Sinclair J.M. Preliminary recommendations on text typology. Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards), June 1996.

58. *Stenström A*-B., Andersen G. More trends in teenage talk: A corpus-based investigation of the discourse items cos and innit // Synchronic corpus linguistics. / Eds C. Percy, C. Meyer, I. Lancashire. – Amsterdam: Rodopi, 1996. – P. 189–203.

59. Svartvik J., Quirk R. A corpus of English Conversation. – Lund: Gleerup, 1980.

60. Svartvik J. Directions in Corpus Linguistics. // Proceedings of Nobel Symposium 92, Stockholm, 4–8 August 1991. – Berlin: Mouton de Cruyter, 1992.

61. Virtanen T. Corpora and discourse analysis // Corpus Linguistics. An International Handbook. Vol. 2. / Eds A. Lüdeling, M. Kytö. – Berlin; N. Y.: Walter de Gruyter, 2008. – P. 1043–1070.

62. Wikipedia – http://en.wikipedia.org/wiki/

Глоссарий

Аболютная частота – количество вхождений слова (словоформы) в данный текст/субкорпус/корпус.

Биграмма – сочетание заданного слова со словом, находящимся справа или слева от него.

Битекст – фрагмент исходного текста и соответствующий ему фрагмент перевода.

Вес текста – условная весовая мера, присваиваемая текстам исследуемого массива на основе их объема.

Выравнивание (стыковка) текстов – установление соответствия фрагментов исходного текста фрагментам перевода, выполняемое вручную или автоматически.

Грамматический разметчик, «тэггер» – программа, выполняющая в автоматическом режиме грамматическую (морфологическую) разметку текстов корпуса.

Графематический анализ – анализ потока символов в текстах на естественном языке, выделение отдельных значимых единиц текста (токенов), возможно, приписывание этим единицах их типов.

Индексирование корпуса текстов – составление в автоматическом режиме списков адресов каждого слова текста, индекса.

Коллокат – слово или словоформа, встречающаяся в качестве ближнего соседа данного слова (словоформы).

Коллокация – регулярное, устойчивое сочетание слов в предложении.

Коллигация - регулярное, устойчивое сочетание слов с учетом морфолого-синтаксических условий, обеспечивающих сочетаемость языковых единиц.

Конкорданс – 1) указатель, связывающий каждое словоупотребление с контекстом; 2) получаемый в автоматическом режиме набор контекстов для заданного явления (слово / словосочетание / грамматическая форма и др.)

Корпус – собрание текстов, обычно в машиночитаемом формате, включающем информацию о ситуации, в которой текст был произведен, такую как информация о говорящем, авторе, адресате или аудитории.

Корпус аннотированный/размеченный – корпус текстов, в котором содержатся специальные метки, позволяющие получать из корпуса данные (статистика, языковые примеры и др.) по каким-либо лингвистическим параметрам (часть речи, грамматическая форма, синтаксическая функция и т.п.).

Корпус выровненный параллельный – параллельный корпус, в котором тексты на одном языке и их переводы на другие языки выровнены по предложениям или по фразам.

Корпус диахронический – корпус текстов, в который включаются тексты, созданные в разные исторические периоды развития языка.

Корпус многоязычный – корпус текстов, включающий в себя текстовые массивы на разных языках.

Корпус мониторный – постоянно пополняемый и обновляемый корпус текстов, создаваемый в целях мониторинга представляемого корпусом подъязыка или языка в целом.

Корпус параллельный – 1) многоязычный корпус, который состоит из текстов на одном языке и их переводов на другой (другие) язык (языки); 2) набор текстов одной и той же тематической области, написанных независимо на двух или нескольких языках.

Корпус полнотекстовый – корпус текстов, состоящий из целых текстов, а не фрагментов.

Корпус сбалансированный – репрезентативный корпус, в котором различные компоненты представлены в «расслоенном» виде, что позволяет создавать схему встречаемости лингвистического явления, исследованного на фоне экстралингвистической информации.

Корпус синхронический/синхронный – корпус текстов, в который включаются только тексты, созданные в течение одного и того же короткого периода времени (например, в течение одного года).

Корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов с использованием компьютерных технологий.

Корпусный менеджер (корпус-менеджер) – специальная информационно-поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации.

Лемма – начальная (словарная) форма для заданной словоформы.

Лемматизация – процесс генерации начальных форм для словоформ.

Нормированная частота – относительная частота, умноженная на миллион; обозначается как ipm (instances per million).

Относительная частота – отношение абсолютной частоты слова (словоформы) к объему корпуса (в словоупотреблениях).

Парсер – компьютерная программа, выполняющая автоматическую разметку текста на синтаксическом или семантическом уровне.

Парсинг – анализ синтаксической структуры предложения и представления ее в виде дерева зависимостей или структуры составляющих.

Разметка корпуса - приписывание текстам корпуса и их компонентам дополнительной информации (метаданных). Метаданные можно поделить на 3 типа: лингвистические, экстралингвистические, данные о структуре текста.

Репрезентативность (представительность, сбалансированность) корпуса текстов – степень представленности в корпусе текстов всех типов текстов, существующих в описываемом языке (подъязыке).

Субкорпус – группа текстов корпуса, объединяемых на основе совпадения какого-либо параметра (язык, жанр и т.п.).

Стыковщик – компьютерная программа, выполняющая в автоматическом или полуавтоматическом режиме выравнивание перевода с исходным текстом.

Токен – конкретное слово в тексте, словоформа, текстоформа, словоупотребление.

Токенизация – разбиение потока символов в текстах на естественном языке на отдельные значимые единицы (токены).

Тэг (тег) – метка, которая присваивается слову или предложению в размеченном корпусе в соответствии с характером разметки.

Утилита – часть пакета программ или отдельная небольшая программа, выполняющая какую-либо операцию по обработке текстов.

Электронный архив – набор текстов в электронном виде, хранящийся на машинном носителе, нестандартизированный и неунифицированный.

Список сокращений

АОТ – Автоматизированная Обработка Текста (название лингвистического Интернет-портала)

ИПС – информационно-поисковая система

МАС – Малый академический словарь

НКРЯ – Национальный корпус русского языка

ОРД – Один речевой день (название проекта)

РГПУ – Российский государственный педагогический университет им. А.И. Герцена

СКАТ – Санкт-Петербургский корпус агиографических текстов

СУБД – система управления базами данных

ЧНК – Чешский национальный корпус

BNC – British National Corpus

COCA – Corpus of Contemporary American English

COLT – [The Bergen] Corpus of London Teenage Language (название корпуса)

CQP – Corpus Query Processor (название корпусного менеджера)

DDC – Dialing-DWDS-Concordance (название корпусного менеджера)

EAGLES – Expert Advisory Group on Language Engineering Standards

ICE – International Corpus of English

KWIC – Key Word In Context

LOB – London-Oslo-Bergen corpus

OED – Oxford English Dictionary

SARA – SGML Aware Retrieval Application

SGML – Standard Generalized Markup Language

SUSANNE – Surface and Underlying Structural Analysis of Naturalistic English (название корпуса)

TACT – Text-Analysis Computing Tools (название программы)

TEI – Text Encoding Initiative

XAIRA – XML Aware Indexing and Retrieval Architecture

WWW – World Wide Web

XCES – XML Corpus Encoding Standard

XML – eXtensible Markup Language

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Приложение

Список некоторых национальных корпусов

Корпус	Название и адрес в сети	Объем (с/у)
Корпус английского языка	Британский национальный корпус http://corpus.byu.edu/ или http://sara.natcorp.ox.ac.uk/	100 млн
Корпус арабского языка	arabiCorpus arabicorpus.byu.edu/	174 млн
Корпус арабского языка	The Quranic Arabic Corpus (Коран) http://corpus.quran.com/ An annotated linguistic resource which shows the Arabic grammar, syntax and morphology for each word in the Holy Quran. The corpus provides three levels of analysis: morphological annotation, a syntactic treebank and a semantic ontology	77 тыс.
Корпус болгарского языка	Болгарский национальный корпус http://www.ibl.bas.bg/BGNC_bg.htm	1,2 млрд
Корпус венгерского языка	Венгерский национальный корпус http://mnsz.nytud.hu/index_hun.html	188 млн
Корпус датского языка	Корпус датского языка KorpusDK http://ordnet.dk/korpusdk	56 млн
Корпус испанского языка	Корпус испанского языка (проект М. Дэвиса) http://www.corpusdelespanol.org/	100 млн
Корпус испанского языка	Corpus de Referencia del Español Actual (CREA) http://corpus.rae.es/creanet.html	150 млн
Корпус итальянского языка	Корпус итальянских текстов Болонского университета CORIS http://corpora.dslo.unibo.it/	130 млн
Корпус китайского языка	The LIVAC Synchronous Corpus (газетный) (Linguistic Variations in Chinese Speech Communities)	450 млн
Корпус китайского	Scripta Sinica database (база данных текстов)	445 млн

языка	http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm	
Корпус немецкого языка	Немецкий справочный корпус Das Deutsche Referenzkorpus – DeReKo) http://www.ids-mannheim.de/kl/projekte/korpora/	5,4 млрд
Корпус немецкого языка	Корпус электронного словаря немецкого языка http://www.dwds.de	2,5 млрд (в поиске 1, 8 млрд)
Корпус немецкого языка	Синтаксически аннотированный корпус немецкого языка NEGRA http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus	355 тыс., (20600 предл.)
Корпус польского языка	PELCRA Reference Corpus of Polish http://pelcra.pl/3-2?lang=pl	100 млн
Корпус польского языка	Корпус польского языка IPI PAN http://korpus.pl/index.php?lang=pl&page=welcome	250 млн
Корпус словацкого языка	Словацкий национальный корпус http://korpus.juls.savba.sk/	1,2 млрд
Корпус словенского языка	Корпус словенского языка http://www.fidaplus.net/	621 млн
Корпус словенского языка	Nova beseda/ http://bos.zrc-sazu.si/a_beseda.html	318 млн
Корпус французского языка	American and French Research on the Treasury of the French Language (ARTFL-FRANTEXT) http://artfl-project.uchicago.edu/content/artfl-frantext	215 млн
Корпус французского языка	Lexiqum http://retour.iro.umontreal.ca/cgi-bin/lexiqum	229 млн
Корпус чешского языка	Чешский национальный корпус http://www.korpus.cz/	1,3 млн
Корпус шведского языка	Банк шведского языка (разные корпусы и словари) http://spraakbanken.gu.se/	1,3 млрд
Корпус японского языка	The Balanced Corpus of Contemporary Written Japanese (BCCWJ) http://www.ninjal.ac.jp/english/products/bccwj/	100 млн

Оглавление

Учебное издание

Виктор Павлович Захаров
Светлана Юрьевна Богданова

КОРПУСНАЯ ЛИНГВИСТИКА

Учебное пособие

Зав. редакцией *Е. П. Парфенова*
Редактор *Н. Г. Михайлова*
Технический редактор *Л. Н. Иванова*
Компьютерная верстка *Н. Г. Михайловой*
Подписано в печать с оригинала-макета 27.05.2013.
Ф-т 60×84/16. Усл. печ. л. 8,60. Уч.-изд. л. 9,62.
Тираж 120 экз. Заказ № .
СПбГУ. РИО. Филологический факультет.
199034, С.-Петербург, Университетская наб., 11.
Типография Издательства СПбГУ.
199061, С.-Петербург, Средний пр., 41.