

---

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

В. П. Захаров, С. Ю. Богданова

# КОРПУСНАЯ ЛИНГВИСТИКА

3-е издание, переработанное



ИЗДАТЕЛЬСТВО САНКТ-ПЕТЕРБУРГСКОГО УНИВЕРСИТЕТА

УДК 81.32  
ББК 81.1-923  
3-38

Авторы:

канд. филол. наук, доцент *В. П. Захаров* (С.-Петербург. гос. ун-т);  
д-р филол. наук, профессор *С. Ю. Богданова* (Иркутский гос. ун-т)

Рецензенты:

д-р филол. наук *С. А. Крылов* (ИВ РАН);  
д-р филол. наук, профессор *Л. Н. Беляева* (РГПУ им. А. И. Герцена);  
канд. филол. наук, доцент *М. В. Хохлова* (СПбГУ)

*Рекомендовано к публикации*

*Учебно-методической комиссией УГСН 45.00.00*

*Языкоизнание и литературоведение*

*Санкт-Петербургского государственного университета*

**Захаров В. П., Богданова С. Ю.**

**3-38** Корпусная лингвистика: учебник. 3-е изд., перераб. — СПб.:  
Изд-во С.-Петербург. ун-та, 2020. — 234 с.  
ISBN 978-5-288-05997-1

Учебник знакомит с концепциями корпусной лингвистики, дает возможность освоить азы корпусных технологий, приобрести навыки работы с корпусами, определить место дисциплины и собственно корпусов в ряду информационных технологий. Базой для создания учебника послужили исследовательская работа и преподавательская деятельность авторов.

Предназначен для студентов, магистрантов и аспирантов филологических и педагогических специальностей, а также для всех интересующихся вопросами корпусной лингвистики.

УДК 81.32  
ББК 81.1-923

ISBN 978-5-288-05997-1

© Санкт-Петербургский  
государственный университет, 2020  
© В. П. Захаров, С. Ю. Богданова, 2020

# Оглавление

Предисловие к третьему изданию.....	7
Предисловие к первому и второму изданиям.....	9
ЧАСТЬ 1. ВВЕДЕНИЕ В КОРПУСНУЮ ЛИНГВИСТИКУ	
<b>Глава 1. Основные понятия корпусной лингвистики .....</b>	<b>11</b>
1.1. Определение корпусной лингвистики .....	—
1.2. Предмет корпусной лингвистики .....	13
1.3. Терминология корпусной лингвистики.....	15
1.4. Направления в лингвистике, предвосхитившие появление корпусной лингвистики .....	17
1.5. Основные характеристики корпусов.....	21
1.5.1. Репрезентативность корпусов .....	—
1.5.2. Прагматическая ориентированность .....	22
1.6. История создания лингвистических корпусов .....	24
<b>Глава 2. Стандартизация в корпусной лингвистике.....</b>	<b>26</b>
2.1. Объекты стандартизации.....	—
2.2. Международные стандарты корпусной лингвистики....	27
2.3. Разметка корпусов в проекте (стандарте) TEI .....	28
<b>Глава 3. Разметка корпусов.....</b>	<b>34</b>
3.1. Понятие разметки .....	—
3.2. Лингвистическая разметка .....	36
3.2.1. Морфологическая разметка .....	37
3.2.1.1. XML формат (формат с ключевыми словами)....	—
3.2.1.2. Позиционный формат кодирования данных разметки.....	40
3.2.1.3. Гибридный формат кодирования данных разметки.....	43
3.2.2. Синтаксическая разметка .....	45
3.2.3. Семантическая разметка .....	50
3.3. Экстралингвистическая разметка.....	54
<b>Глава 4. Типология корпусов.....</b>	<b>56</b>
4.1. Классификация корпусов по различным основаниям ..	—
4.2. Особенности корпусов отдельных типов.....	61

---

## Оглавление

---

4.2.1. Параллельные корпусы.....	61
4.2.2. Корпусы устной речи .....	64
4.2.3. Учебные корпусы текстов .....	67
Вопросы и задания для самоконтроля .....	69
 ЧАСТЬ 2. СОЗДАНИЕ КОРПУСОВ	
<b>Глава 5. Традиционная технология создания корпусов .....</b>	70
5.1. Проектирование и технологический процесс создания корпусов.....	—
5.2. Отбор источников. Критерии отбора .....	72
5.3. Основные процедуры обработки входных текстов .....	74
5.4. Как создать собственный корпус?.....	77
<b>Глава 6. Создание корпусов на базе веба.....</b>	79
6.1. Поисковые системы Интернета как корпусы.....	—
6.2. Веб как корпус.....	80
6.3. Технология WaC.....	83
<b>Глава 7. Обзор существующих корпусов различных типов.....</b>	85
7.1. Зарубежные корпусы.....	—
7.2. Корпусы русского языка.....	95
7.2.1. Первые корпусы русского языка .....	—
7.2.2. Современные корпусы русского языка .....	99
7.2.2.1. Национальный корпус русского языка .....	—
7.2.2.2. Хельсинкский аннотированный корпус (ХАНКО).....	101
7.2.2.3. Корпусы университета г. Лидс.....	102
7.2.2.4. Другие текстовые корпусы русского языка .....	103
7.2.2.5. Устные корпусы русского языка.....	—
7.2.2.6. Мультимедийные корпусы русского языка .....	105
7.3. Специальные корпусы .....	107
Вопросы и задания для самоконтроля .....	109
 ЧАСТЬ 3. ПОЛЬЗОВАНИЕ КОРПУСАМИ	
<b>Глава 8. Корпусные менеджеры.....</b>	110
8.1. Корпус как поисковая система.....	—
8.2. Функциональные возможности корпусных менеджеров.....	115

---

8.3. Языки запросов корпусных менеджеров.....	116
8.4. Язык запросов корпусного менеджера Sketch Engine ....	118
8.5. Язык регулярных выражений RegEx .....	121
8.6. Сервисные функции .....	127
<b>Глава 9. Способы использования корпусов .....</b>	<b>132</b>
9.1. Пользователи корпусов.....	—
9.2. Что можно получить из корпуса?.....	133
9.2.1. Эмпирическая поддержка .....	—
9.2.2. Статистическая информация.....	135
9.2.3. Метаинформация.....	135
Вопросы и задания для самоконтроля .....	—
<b>ЧАСТЬ 4. ЛИНГВИСТИЧЕСКИЕ ИССЛЕДОВАНИЯ НА БАЗЕ КОРПУСОВ</b>	
<b>Глава 10. Лексикографические исследования, основанные на корпусах.....</b>	<b>137</b>
10.1. Пример одного лексикографического исследования...	138
10.1.1. Распределение <i>deal</i> по регистрам.....	140
10.1.2. Распределение смыслов (значений) по регистрам .....	143
10.1.3. Слово <i>deal</i> как глагол .....	148
10.2. Анализ использования слов, кажущихся синонимами .....	149
10.2.1. Распределение по регистрам синонимичных английских прилагательных <i>big</i> , <i>large</i> и <i>great</i> .....	149
10.2.2. Удаленные коллокаты <i>large</i> .....	156
<b>Глава 11. Грамматические исследования, основанные на корпусах .....</b>	<b>158</b>
11.1. Распределение и функции номинализаций .....	159
11.1.1. Анализ распределения номинализаций по регистрам .....	—
11.1.2. Распределение и функция суффиксов номинализаций .....	161
11.2. Распределение грамматических категорий .....	163
11.2.1. Частотность грамматических категорий .....	164
11.2.2. Сравнение соотношения «существительное/ глагол» по регистрам .....	166

<b>Глава 12. Исследования дискурса, основанные на корпусах.....</b>	<b>167</b>
12.1. Характеристики референциальных выражений .....	169
12.1.1. Распределение референциальных выражений по регистрам .....	169
12.1.2. Техника интерактивного анализа: кодирование характеристик референциальных выражений....	173
12.2. Распределение обращений в неформальной беседе...	175
12.3. Пример исследования дискурса на материале речевого корпуса.....	176
<b>Глава 13. Корпусные методы исследования .....</b>	<b>179</b>
13.1. Применение корпусных методов сбора, обработки и аннотирования текстового материала .....	180
13.1.1. Корпусы делового языка.....	—
13.1.2. Корпусы диалектов.....	182
13.1.3. Корпус устной речи «Один речевой день» .....	183
13.1.4. Учебный прагматический корпус.....	185
13.2. Применение корпусных методов извлечения информации из русскоязычных корпусов текстов ....	186
13.2.1. Корпусы и переводная лексикография .....	—
13.2.2. Веб-корпусы: <i>pro et contra</i> .....	190
13.3. Применение статистических методов в корпусных исследованиях.....	193
13.3.1. Корпусный анализ фразеологии.....	194
13.3.2. Диахронические исследования грамматики .....	198
13.4. Выделение коллокаций статистическими методами ...	200
Вопросы и задания для самоконтроля .....	204
<b>Заключение .....</b>	<b>205</b>
Темы докладов, рефератов, курсовых работ .....	207
Рекомендуемая литература.....	211
Список цитируемых источников .....	214
Глоссарий .....	226
Список сокращений .....	230
Предметный указатель.....	231

# Предисловие к третьему изданию

Предлагаемый учебник является результатом научной и педагогической деятельности авторов, а также обобщением многочисленных материалов по корпусной лингвистике, опубликованных в России и за рубежом, естественно, малой их части. На его основе построены лекционные курсы по корпусной лингвистике и смежным с ней дисциплинам, читаемые на протяжении многих лет В. П. Захаровым в Санкт-Петербургском государственном университете и С. Ю. Богдановой в Иркутском государственном университете. Материал, представленный в учебнике, также может быть использован в курсах лекций по дисциплинам «Информационные и коммуникационные технологии в науке и образовании», «Основы прикладной лингвистики», «Кvantитативная лингвистика», «Корпусы при автоматической обработке текста», «Компьютерные методы в лингвистических исследованиях», «Корпусы и переводоведение» и др.

По сравнению со вторым изданием главные изменения следующие:

- переработаны многие прежние и добавлены новые разделы, в частности раздел 5.4. «Как создать собственный корпус?», глава 6 «Создание корпусов на базе веба», глава 13 «Корпусные методы исследования» и др.;
- добавлена или исправлена информация о корпусах, существовавших на момент подготовки второго издания, и новых;
- добавлена информация о новых корпусных инструментах, появившихся или претерпевших изменения после выхода второго издания;
- отражены некоторые новые публикации;
- изменена структура учебника.

---

## Предисловие к третьему изданию

---

В данном издании учебник состоит из 13 глав, разбитых на 4 части: «Введение в корпусную лингвистику», «Создание корпусов», «Пользование корпусами» и «Лингвистические исследования на базе корпусов».

Современное развитие лингвистики как эмпирической науки диктует необходимость использования новых, объективных методов исследования. Корпусная лингвистика является одним из разделов науки о языке, который предоставляет такие возможности. Какими воспользоваться — об этом авторы постарались рассказать в учебнике.

# Предисловие к первому и второму изданиям

Предлагаемый вашему вниманию учебник является своего рода обобщением многочисленных разрозненных материалов, опубликованных за последние два десятилетия в России и за рубежом. Данные материалы 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 пособии, также может 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 учебного пособия:

- 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.

---

## Предисловие к первому и второму изданиям

---

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, целями и задачами, типами существующих корпусов. Вторая часть — «Создание корпусов» — описывает в общих чертах технологические процессы, связанные с проектированием корпусов, 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

### Глава 1. Основные понятия корпусной лингвистики

#### 1.1. Определение 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* (или обычно просто *корпусом текстов*) понимается большой, представленный в машиночитаемом формате, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач. Имея в виду круг задач (подчас достаточно широкий), для решения которых создается тот или иной корпус, можно говорить, что корпус всегда прагматически ориентирован.

В настоящее время существует множество определений понятия «корпус». Например, определение, приведенное в учебнике Э. Файнегана, гласит: *корпус* — презентативное собрание текстов, обычно в машиночитаемом формате, включающее информацию о ситуации, в которой текст был произведен, такую как информация о говорящем, авторе, адресате или аудитории [Finegan, 2004].

Википедия определяет корпусы как большие и структурированные наборы текстов (теперь обычно в электронном виде), которые используются для статистического анализа и проверки гипотез, подтверждения или обоснования лингвистических правил.

Т. Мак-Энери и Э. Вилсон дают следующее определение: *корпус* — это собрание языковых фрагментов, отобранных в соответствии с четкими языковыми критериями для использования в качестве модели языка [McEnergy, Wilson, 2001].

В приведенных определениях подчеркиваются основные черты современного корпуса текстов: цель («логическая идея», прагмати-

ческая ориентация), машиночитаемый формат, репрезентативность как результат особой процедуры отбора текстов, наличие металингвистической информации. Стандартизованное представление словесного материала на машинном носителе позволяет применять стандартные программы его обработки.

Целесообразность создания и смысл использования корпусов определяются следующими предпосылками:

- достаточно большой (репрезентативный) и сбалансированный объем корпуса гарантирует типичность данных и обеспечивает полноту представления всего спектра языковых явлений;
- данные разного типа находятся в корпусе в своей естественной контекстной форме, что создает возможность их всестороннего и объективного изучения;
- 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 называют *корпусным менеджером* (или корпус-менеджером) (англ. *corpus manager*). Это специализированная поисковая система, включающая в себя программные средства для поиска запрашиваемых данных в корпусе и предоставления их пользователю в удобной форме, а также для получения статистической информации.

Поиск в корпусе позволяет по любому слову построить *конкорданс* — список всех употреблений данного слова в контексте со ссылками на источник.

Однако кроме этого корпусы могут использоваться для получения справок о характеристиках текста или лексических единиц, статистических данных о языковых единицах и о лингвистических категориях и метаданных (частоте словоформ, лексем, грамматических категориях, изменении частот и контекстов в различные периоды времени), данных о совместной встречаемости лексических единиц, жанрово-стилистических характеристиках и т. п. Эти статистические данные могут выдаваться непосредственно (например, частотный список), а могут использоваться для «внутренних» подсчетов и выдачи новых данных, непосредственно в корпусе не заложенных, например количественное выражение устойчивости соче-

таний в тексте, парадигматическая (семантическая) кластеризация лексических единиц, выявление ключевых слов текста.

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.

Лингвистов-исследователей все больше интересуют функции дополнительной, можно сказать, интеллектуальной обработки корпусных данных. И такие программы есть, они представляют собой уже не просто корпусный менеджер как информационно-поисковую систему фактографического типа, а сложный конгломерат программных, лингвистических, математических средств, обеспечивающий широкий набор разнообразных лингвистических функций. Мы предлагаем для этого понятия название «корпусная служба».

## 1.2. Предмет корпусной лингвистики

Сегодня корпусная лингвистика часто понимается как новая лингвистическая дисциплина, которая связана 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, а именно размеченные тексты; 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 как языковых исследований, так и не только языковых. Корпусные методы лежат в основе новой дисциплины, которая получила название «культурометрия» (*culturomics*) и распространяется на все области гуманитарных исследований.

Корпусную лингвистику называют «пучком методов из разных областей лингвистических исследований» [Lüdeling, Kytö, 2008]. 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 [Гвишиани, 2008]. Многие виды лингвистического анализа наилучшим образом развиваются на прочной и обширной базе эмпирических данных.

Задаваясь вопросом о месте корпусной лингвистики в лингвистике вообще, видимо, правильнее всего будет сказать, что это методология лингвистического исследования, применимая практически к любой области лингвистики. Однако существует и другой взгляд: корпусная лингвистика — это, собственно, и есть настоящая научная лингвистика. В англоязычной литературе эти подходы — корпусная лингвистика как методология лингвистики и как отдельная наука — получили название *corpus-based* (корпусно-ориентированный подход) и *corpus-driven* (корпусно-управляемый подход).

Первый подход предполагает, что корпусы используются для проверки лингвистических теорий или гипотез, чтобы их подкрепить, подтвердить, опровергнуть или уточнить. Второй подход провозглашает, что корпус сам является главным и единственным источником наших теорий о языке, корпусная лингвистика получает здесь статус теории [Tognini-Bonelli, 2001, р. 1] и рассматривается как «важнейший концепт в лингвистической теории» [Stubbs, 1993, р. 24]. Это значит, что корпус неявно содержит в себе теорию языка и нужно ее оттуда только «добыть» [Sinclair, 2004, р. 191]. «Теория не существует независимо от данных» [Tognini-Bonelli, 2001, р. 84–85]. Это понимание возвращает нас к работам американских структуралистов первой трети XX в.

В недрах корпусной лингвистики этот подход называют неоферсианским (*neo-Firthian*), так как он сильно связан с понятием коллокации, введенным Дж. Р. Фёрсом (Firth). Может быть, самой знаменитой цитатой в корпусной лингвистике является высказывание Дж. Р. Фёрса: «Вы поймете слово по его окружению» (“You shall know a word by the company it keeps”) [Firth, 1957, р. 11]. Суть этого подхода заключается в том, что значение слова (равно как и другие лингвистические концепты) существует только в контексте (в тексте). Предполагается, что аналитик, исследующий данные, не использует никаких априори установленных теоретических концепций. Другой краеугольный камень подхода неоферсианцев к изучению языка — это понятие дискурса. Дискурс для них — это не только текст, «практика» языка, но и способ реализации самого языка или подъязыка, не только способ говорения, но и способ мышления. И здесь воззрения ученых, исповедующих это направление и использующих

корпусные ресурсы и методы, по ряду позиций стыкуются с психолингвистикой и с социолингвистикой [Sinclair, 2004].

Можно также привести высказывание В. А. Плунгяна (лекция в Европейском университете в Санкт-Петербурге) о том, что если раньше лингвистика стояла на двух «китах» — лексике и грамматике, то теперь к ним добавилась третья ипостась — корпус.

На практике оба вышеупомянутых подхода имеют много общего. И можно отметить, что многие публикации лингвистов, исповедующих корпусно-управляемый подход, на самом деле представляют собой корпусно-ориентированные исследования.

На наш взгляд, важным аспектом в определении корпусной лингвистики является то, что это не просто методология исследования языка — это наука, в недрах которой формируется сам объект исследования. Э. Финеган определяет корпусную лингвистику как деятельность, требующуюся для составления и использования корпуса и направленную на исследование естественного употребления языка [Finegan, 2004]. В этом определении подчеркивается созидательная направленность корпусной лингвистики. Ее двойственный характер (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, сам 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.

### 1.3. Терминология корпусной лингвистики

Говоря о терминологии в области корпусной лингвистики, прежде всего следует сослаться на обширный труд «A glossary of corpus Linguistics» [Baker, McEnergy, Hardie, 2006].

В русском языке терминология корпусной лингвистики пока окончательно не установилась в силу ряда причин: зарождение корпусной лингвистики в США и Великобритании и ее более позднее развитие в России обусловили тот факт, что терминоло-

гия складывалась и продолжает развиваться в недрах английского языка. Русские термины в основном представляют собой английские заимствования, некоторые из них, но 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 «корпusa». Для значения «массив», которое имеет место в случае языковых корпусов, именительный падеж множественного числа должен быть «кóрпусы», и, соответственно, прилагательное «кóрпусный» должно произноситься с ударением на первом слоге [Большой толковый словарь русского языка, 1998]. В то же время наблюдение над узлом специалистов пока свидетельствует в пользу форм «корпusa», «корпуснóй», «корпуснáя», которые используются часто, так что можно, видимо, с осторожностью сказать, что в настояще время этот вопрос остается открытым. Правила, регламентирующего употребление той или иной формы применительно к корпусной лингвистике, пока нет, хотя, как представляется, «победить» должен вариант «кóрпусы». В данном учебнике авторы будут использовать именно этот вариант.

Имеется проблема с дефинициями и других терминов, когда они используются в публикациях или в документации по корпусной лингвистике как общепринятые в лингвистике (слово, словоформа, биграмма, коллокация, метаданные), так и как специальные (*ipm*, корпус-менеджер, токен, коллигация и др.).

В данный момент терминология корпусной лингвистики частично отражена в глоссариях к учебникам [Грудева, 2017; Щипицына, 2015; Копотев, 2014], а также в тезаурусе по компьютерной лингвистике (<https://uniserv.iis.nsk.su/thes/>). Наш вклад в развитие корпусной терминологии мы попытались отразить в глоссарии (см. приложение).

#### **1.4. Направления в лингвистике, предвосхитившие появление корпусной лингвистики**

В первой половине 1990-х годов корпусная лингвистика окончательно сформировалась как новое лингвистическое направление. Но правильно ли будет сказать *новое*?

Знаменитая ключевая фраза мольеровского героя из пьесы «Мещанин во дворянстве» звучит так: «...я и не подозревал, что вот уже более сорока лет я говорю прозой». Открытие, сделанное господином Журденом, должно изобличать, конечно, его безграмотность, однако можно сказать, что мы действительно говорим прозой. Так же можно сказать, что лингвисты всю жизнь занимались корпусной лингвистикой, не подозревая об этом. Не случайно одна из статей В. Н. Фрэнсиса (W. N. Francis) называется «Согорга В. С.» (*before Christ* — до рождества Христова). Здесь, конечно, игра слов: автор имел в виду *корпусы до компьютеров* (*Corpora Before Computers*). В статье речь идет о том, что идеи корпусной лингвистики действительно зародились задолго до компьютерной эры [Francis, 1992]. Лингвисты и лексикографы уже давно в своей работе используют эмпирический материал, цитаты из текстов, которые выписывались на карточки и образовывали «корпусы» под названием «карточки».

Основной выходной продукт корпусов — это конкорданс, но и он «изобретен» давным-давно. Первая «конкорданция» появилась в начале XIII в. («Concordantiae morales sacrae scripturae» — «Нравственная конкорданция Священного Писания»). Это был своего рода предметный конкорданс к текстам Библии. Вслед за ней около 1230 г. появилась конкорданция к «Вульгате» Гуго де Сен-Шера, первого кардинала доминиканского монастыря святого Иакова в Париже (Concordantiae Sancti Jacobi). Для ее составления автор воспользовался услугами 500 доминиканцев, собратий своего монастыря. При цитируемых словах были даны подтверждения из Библии с указанием места, откуда они взяты.

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 корениются в традиции конца XVIII — XIX в., 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» всей современной лингвистики, включая все вышеописанные направления [Lüdeling, Kytö, 2008].

**1. Историческая лингвистика: изменения в языке и реконструкция (сравнительно-исторический метод).** Одно из главных направлений, повлиявших на современную корпусную лингвистику, пришло из сравнительно-исторического языкоznания. Это неудивительно, поскольку лингвисты, занимающиеся историческими исследованиями, всегда использовали тексты или собрания текстов как основные свидетельства. Многие технологии, развитые в XIX в., и в настоящее время используются для реконструкции более древних языков (праязыков) или установления связей между языками. В индоевропейской традиции изучение языковых изменений и попытки реконструкции зависели от ранних текстов или корпусов (исторических памятников). Я. Гримм и позднее младограмматики поддерживали 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 исторические корпусы.

Появление огромного количества текстов, доступных в электронном формате, предоставило лингвистам возможность широко применять в лингвистическом анализе статистические методы, 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 классиков для иллюстрации каждого своего положения в области фонологии, морфологии и синтаксиса. Сегодня составители грамматик также используют корпусный подход, теперь корпусы включают не только классику, но и другие типы текстов и позволяют описать язык более адекватно. В частности, большой интерес сейчас вызывает грамматика устной речи.

В грамматических описаниях языка корпусы можно использовать для получения информации о частотных характеристиках использования разных вариантов, регистров (жанров)<sup>1</sup> и т. п.

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 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 помощников 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), в котором искомое слово или конструкция выделяются в центре рабочего поля, а справа и слева отображается контекст. Компьютеры облегчили поиск и классификацию примеров, но идеи использования текстов из корпуса все еще очень схожи с теми, которых придерживались ранние лексикографы и филологи. И как мы уже писали выше, не только филологи.

---

<sup>1</sup> 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., например: Lee D. Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle // Language Learning & Technology. 2001. Vol. 5, No. 3. 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, а «регистры» акцентируют внимание на параметрах ситуации языкового употребления и имеют естественную ассоциацию с определенными *лингвистическими чертами*). Это позволяет рассматривать устную речь (*spoken*) наравне с академической прозой (*academic*) и художественной литературой (*fiction*) как регистр..

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 общении. В этом отношении корпусы как источники эмпирических данных играют важную роль в лингводидактике. При обучении языку корпусы представляют собой источник, призванный пробудить у студентов интерес и вовлечь их в самостоятельное изучение аутентичного языкового использования. Важное применение корпусных данных — технологии Computer-Assisted Language Learning (CALL), где основанное на корпусе программное обеспечение используется для поддержки интерактивной учебной деятельности, осуществляющейся студентами при помощи компьютера [Потапова, 2005].

**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, за исключением существенной отличительной черты: 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 (диалектов, социолектов, регистров). Математические методы (например, мультифакторный анализ, то есть 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 и развитием методов искусственного интеллекта, базирующихся на больших данных (Big Data).

## 1.5. Основные характеристики корпусов

### 1.5.1. Репрезентативность корпусов

Термин «корпус» обычно обозначает собрание текстов конечно-го фиксированного размера. С течением времени объем и состав корпуса могут меняться, однако эти изменения должны либо не менять его структуру, либо менять ее обоснованно. Представительность корпуса получила название **репрезентативность** (англ. *representativeness*).

Почти в любом случае проблема репрезентативности существует, и ее надо решать. Допустим, при создании корпуса писем Н. В. Гоголя очевидно, что туда должны войти все письма Н. В. Гоголя. Но это, скорее, исключение. В реальном языке или подъязыке текстов, как правило, гораздо больше. Тогда встают две проблемы — объема и отбора.

Проблема объема явно была сформулирована в 1960–1970-е годы XX в. при создании частотных словарей, когда обсуждалось понятие представительной выборки — такого количества языкового материала, после достижения которого относительные частоты языковых единиц практически не меняются. Объем первых корпусов составлял 1 млн словоупотреблений (Брауновский корпус, корпус Ланкастер-Осло-Берген, корпус Частотного словаря русского языка под ред. Л. Н. Засориной). Такой объем не позволял отразить язык во всем его многообразии. Затем стали считать, что общезыковой (национальный) корпус должен включать не менее 100 млн словоупотреблений. Очевидно, что для изучения многих языковых явлений и этого объема недостаточно. Поэтому сейчас создаются корпусы, где счет словоупотреблениям идет на миллиарды.

Однако статистические критерии оценки не всегда являются единственными или определяющими, поскольку корпус выступает как некоторый объект, призванный послужить *моделью* внешней по отношению к нему реальности.

Вторая проблема формирования представительной выборки — проблема отбора, призванная ответить на вопрос о том, из каких текстов сформировать тот самый минимально необходимый объем. Поэтому появилось еще одно важное понятие корпусной лингвистики — **сбалансированность** корпуса (англ. *balance*). Этую характеристику обеспечить еще труднее, особенно применительно к националь-

ным корпусам. Национальный корпус представляет данный язык на определенном этапе (или этапах) его существования во всем многообразии жанров, стилей, территориальных и социальных вариантов, временных периодов и т. п. Если корпус — это уменьшенная модель языка (или подъязыка), то в нем пропорционально должны быть представлены тексты, относящиеся к разным подмножествам языка (подъязыка). Можно сказать, что применительно к языку в целом мы этих пропорций не знаем. Тем не менее к этому нужно стремиться как на этапе проектирования корпуса, так и на этапе его развития.

Можно сказать, что *репрезентативность и сбалансированность* корпуса обеспечивают достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов, то есть способность отражать все свойства языка или подъязыка. И именно они определяют достоверность полученных на материале корпуса результатов.

### 1.5.2. Прагматическая ориентированность

Практика разработки и применения электронных корпусов текстов показала, что невозможно создать универсальный корпус, обеспечивающий решение всех задач. Задачи и цели любого исследования определяют тип корпуса, правила отбора текстов, а также способ и степень их обработки. Корпусы всегда создаются для решения определенной задачи или круга задач. Это определяет как наполнение корпуса текстами (например русская драма XIX в., тексты языка охотников), так и разметку корпуса.

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

- корпусы первого типа — универсальные, отражающие все многообразие речевой деятельности;
- корпусы второго типа — специфичные, отражающие бытование некоторого языкового или культурного явления в общественной речевой практике, например корпус пословиц или корпус политических метафор в газетной речи;
- корпусы третьего типа — специфичные, создаваемые для решения специальной задачи, например для обучения, для задач социолингвистики, для отладки систем машинного перевода.

Речевая действительность чрезвычайно разнообразна, и разнобразие зафиксированных в ней лингвистических явлений просто необозримо. В 1960-е годы корпусы текстов, относящиеся к *первому типу*, претендовали на то, что они суть универсальные, то есть отражают статистически корректно всю картину бытования данного языка или некоторый представительный ее фрагмент [McEnergy, Wilson, 2001]. Например, *Брауновский корпус* текстов (The Brown Corpus<sup>2</sup>) был создан для отражения письменной речи США начала 1960-х годов с удовлетворительной для того времени степенью репрезентативности. Отобранные тексты представляли 15 жанров (registров), из которых было сделано от 6 до 80 выборок. Как мы сейчас понимаем, и репрезентативность, и сбалансированность этого корпуса довольно спорны.

В корпусах *второго типа* критерием репрезентативности будет служить требование максимально полного и объективного представления об интересующем его создателей явлении. Так, можно сказать, что очень большой корпус английских пословиц максимально репрезентативно отражает этот лингвистический объект в английском языке определенного времени и географического региона.

Наполнение корпусов *третьего типа* определяется спецификой той задачи, для решения которой они создаются.

Методология конструирования такого объекта, как корпус, должна зависеть от его типа. Эта проблема является актуальной и недостаточно проработанной. Методология построения корпусов первого типа так или иначе основывана 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 типов должна корректно отражать частные, единичные языковые явления в корпусе текстов, специально созданном для их отражения. Теория и практика показывают, что оба эти подхода часто применяются в комбинированном виде.

Корпусная pragmatика включает в себя и аудиторию, для которой создается корпус, и тип задач.

---

<sup>2</sup> Полное название корпуса — The Brown Standard Corpus of American English. Он был разработан в Брауновском университете (Brown University) в США в 1963 г.

## 1.6. История создания лингвистических корпусов

Лингвисты собрали первые корпусы компьютеризированных текстов в 1960-е годы. Первый компьютеризированный корпус — Брауновский — 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). Репрезентативность и сбалансированность корпуса обеспечивали следующие пятнадцать жанров:

- 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.

Авторы корпуса В.Н.Френсис (W.N.Francis) и Г.Кучера (H.Kucera) сопроводили его большим количеством материалов первичной статистической обработки — частотным и алфавитно-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, или кратко LOB) по названиям британского и двух норвежских университетов.

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 (табл. 1.1) [Johansson, 2008].

*Таблица 1.1. 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

Очевидно, что в последующие годы количество и многообразие создаваемых корпусов только увеличивались. Среди современных корпусов *английского языка* (как британского, так и американского)

го 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 английского* (Corpus of Contemporary American English — COCA) и др. В настоящее время корпусы созданы для многих языков мира (см. гл. 7).

Ключевым моментом развития корпусной лингвистики в России стало создание Национального корпуса русского языка (2004) (<http://ruscorpora.ru>). Однако еще в Советском Союзе предпринимались попытки создания корпусов. См. об этом в п. 7.2.1.

В первой половине 1990-х годов 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 [Svartvik, 1992].

И все же будет правильным сказать, что корпусная лингвистика — это наука XXI в. Сегодня это неотъемлемая часть лингвистики, можно сказать, ее «тело», а «двигателем» является компьютерная лингвистика. Корпусная лингвистика тесно взаимодействует с компьютерной лингвистикой, используя ее достижения и, в свою очередь, обогащая ее.

## Глава 2. Стандартизация в корпусной лингвистике

### 2.1. Объекты стандартизации

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 унифицируют собственно разметку корпусов. Иногда их называют стандартами кодирования. Также важным является вопрос, связанный со сравнением разных корпусов, в том числе с оценками их пригодности к различным заданиям. Их называют «стандартами оценки».

Особую сложность представляет стандартизация транскрибирования устной речи и исторических корпусов. Хотя в области графической фиксации устной речи даже при отсутствии единого и обязательного для всех стандарта достигнут некоторый прогресс

(связанный прежде всего с наличием прецедентов), то в описании невербальной составляющей естественно-языковой коммуникации стандарты до сих пор не выработаны, что затрудняет дальнейшее продвижение в этом направлении [Баранов, 2007].

С точки зрения стандартизации оценки, 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.

Во многих случаях единые форматы представления данных позволяют 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.

## 2.2. 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, <http://www.ilc.cnr.it/EAGLES/isle/right.html>) и на рекомендациях EAGLES (Expert Advisory Group on Language Engineering Standards, <http://www.ilc.cnr.it/EAGLES/home.html>). Среди них в первую очередь следует назвать форматы CDIF (Corpus Document Interchange Format, [www.natcorp.ox.ac.uk/archive/vault/tgcw30.pdf](http://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/>). В настоящее время эти и другие стандарты «собираются» и обобщаются под эгидой комитета Международной организации по стандартизации ISO/TC 37. Стандарты ISO 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.

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.

### **2.3. Разметка корпусов в проекте (стандарте) TEI**

Наиболее проработанными являются рекомендации проекта Text Encoding Initiative (TEI). Начало проекта по созданию системы кодирования текстов связано с семинаром в Бассарском колледже в 1987 г., где присутствовали представители текстовых архивов, научных обществ и исследовательских центров. Участники семинара обсуждали возможность создания стандартной схемы кодирования текстовых документов. Проект TEI стартовал в 1988 г. Основные понятия и структура TEI практически не менялись на протяжении более десяти лет. Четвертая версия TEI (TEI P4) (2002 г.) характеризуется дополнениями, связанными с внедрением языка XML. TEI P5, последняя версия рекомендаций, была опубликована в 2007 г. и постоянно обновляется. Последний релиз 3.4.0 был выпущен 23 июля 2018 г. (<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/>). Информацию о проекте в целом и современном состоянии разработки,

а также все предыдущие релизы можно найти на сайте проекта (<http://www.tei-c.org/guidelines/p5/>).

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 самых разных типов, предоставляет возможность описания и идентификации других форматов данных, например графических и звуковых материалов. Главная цель проекта — разработка форматов для обмена данными в гуманитарной области.

Рекомендации TEI призваны:

- определить единый синтаксис формата;
- определить метаязык для описания схем представления и кодирования данных;
- описать существующие схемы кодирования на 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 рекомендаций без специального программного обеспечения.

TEI поддерживают международные организации: 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 используется язык XML (ранее SGML), позволяющий 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, то есть единые цельности, состоящие из множества текстов*.

Это объясняется тем, что, хотя каждый отдельный фрагмент текста в корпусе имеет право считаться самостоятельным, в научных целях каждый фрагмент рассматривается и как составляющая большего объекта. Корпусы и другие типы составных текстов (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, относящихся к информации любого вида, рекомендации TEI предлагают 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-либо причине рассматриваются как единое целое, например тексты одного автора, стихотворный цикл и т. п.

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 употребляется термин *участник речевого акта*, 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* (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 фрагменте 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 связанным нетегированным текстом (как описание оформления сцены перед началом спектакля) или 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 собственное) — 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. п.;

**<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 «контекстуальная информация». 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 характеристики, размах выборки равномерно представлен во всем корпусе или представлен пропорционально численности фрагментов, взятых для составления корпуса). Метаинформация 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 для удобства выбора подмножества корпуса по определенным признакам.

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

**id** — уникальный идентификатор документа в корпусе;

**target** — имя файла, в котором находится документ;

**type='text'** — тип описания, для текстовых корпусов всегда **«text»;**

**lang** — язык, на котором написан документ, для корпусов на русском языке значение атрибута равняется **«ru»**, в TEI используется указание языка по стандарту ISO 639 (атрибут *lang* обычно задает значение *по умолчанию*). Это значение может быть переопределено для отдельного предложения или слова, если, например, в русский текст включен фрагмент на другом языке; в TEI предусмотрен также тег **<foreign>** для иноязычных 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>** полезен тег **<profileDesc>**, который содержит информацию об общем классе текстов, например художественной литературе, публицистике, устной речи 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 обычно принято считать размер документа в словах. Должны быть сформулированы правила для подсчета слов, например, можно считать словом последовательность символов от пробела до пробела, можно, наоборот, только последовательности букв из кириллицы/латиницы, можно только из кириллицы, можно считать многословные единицы, например *так как*, *друг друга*, за одно слово. Учитывая, что размер — одна из главных характеристик корпуса, требуется точное указание единицы измерения для этого параметра. Обычно длина корпуса указывается в токенах);

**<sponsor>** — элемент, в котором мы можем сослаться на соответствующего спонсора (в TEI имеется еще синонимичный элемент **<funder>**);

**<respStmt>** — информация о человеке/людах, внесших интеллектуальный вклад в создание данного электронного доку-

мента (не авторы и не спонсоры); **<respStmt>** задает информацию с помощью элементов **<name>** и **<resp>** для указания природы интеллектуального вклада.

## Глава 3. Разметка корпусов

### 3.1. Понятие разметки

Разметка корпуса (*tagging, annotation*) заключается в присыпывании текстам и их компонентам специальных тегов: собственно лингвистических, описывающих лексические, грамматические и прочие характеристики элементов текста, и внешних, экстралингвистических (сведений об авторе и о тексте: автор, название, год и место издания, жанр, тематика). Особое значение имеет лингвистическая разметка.

Разметка корпусов представляет собой трудоемкую операцию, учитывая огромные размеры современных корпусов. Если для некоторых видов разметки, в частности анафорической и просодической, создание автоматических систем пока представляется довольно сложным и основная часть работы проводится вручную, то для морфологического и синтаксического анализа и, соответственно, разметки существуют различные программные средства, которые принято называть соответственно *теггеры* (*taggers*) и *парсеры* (*parsers*).

В результате работы программ автоматической *морфологической разметки* каждой лексической единице присыпаются грамматические характеристики, включая признак части речи, лемму (первоначальную форму слова, зафиксированную в словаре), набор граммем (например род, число, падеж, одушевленность/неодушевленность, переходность/непереходность и т. п.).

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

Однако автоматический анализ естественного языка 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. Важный момент — это снятие неоднозначности (*disambiguity*), морфологиче-

ской, синтаксической, семантической и т. п. 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 др. Несмотря на наличие программ автоматического снятия неоднозначности (*disambiguation*), здесь также требуется участие лингвиста.

Снятие неоднозначности (иначе говоря, разрешение омонимии) в целом является одной из важнейших и сложнейших задач компьютерной лингвистики. При создании корпусов для снятия неоднозначности используются автоматический и ручной<sup>3</sup> способы обработки или их комбинация.

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

Автоматический анализ, кроме того, не безошибочен, причем только морфологически размеченные корпусы можно использовать в работе без ручной корректировки. Качество разметки во многом зависит от типа текста. Так, в веб-корпусах относительное количество неправильно приписанных лемм и грамматических признаков выше, чем в обычных корпусах.

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 подходит значение части речи «существительное», поскольку в английском языке прилагательные обычно предшествуют существительным. Тогда *deal* в *a good deal of trouble* должно быть помечено как существительное.

---

<sup>3</sup> Здесь и далее под ручной обработкой понимается дополнительный интеллектуальный анализ с привлечением других методов анализа.

### 3.2. Лингвистическая разметка

Под *лингвистической разметкой* подразумевается любая аннотация, основанная на лингвистических характеристиках текста. Языковой корпус, как правило, содержит в своем описании аналитико-лингвистические параметры, используемые в различных исследованиях. Существует несколько механизмов, с помощью которых можно представить практически любой тип разметки в стандартном виде или по шаблону, индивидуальному для данного документа.

Среди лингвистических типов разметки выделяются морфологическая, синтаксическая, семантическая, анафорическая, просодическая, дискурсная и др. При их осуществлении соблюдаются основные принципы:

- теоретически нейтральная (традиционная) схема разметки;
- общепринятая система лингвистических понятий;
- мотивированность введения параметров;
- следование международным стандартам.

Данные лингвистической разметки можно добавлять к текстовым элементам разных уровней. Например, код класса слов или код частеречной принадлежности может быть привязан к каждому слову (токену) или группе токенов, которая может быть неразрывной или разрывной. Синтаксический код может быть закреплен за предложением или за синтаксическим отношением.

С точки зрения технологии, разметка может быть автоматической, ручной или автоматической с ручной правкой. Среди специальных программ для создания корпусов особое место занимают программы автоматической разметки.

#### 3.2.1. 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, в-третьих, успехи в компьютерной морфологии позволяют автоматически с большой степенью правильности размечать корпусы больших размеров.

В 1980 г. появилась размеченная версия Брауновского корпуса. Морфологическая разметка фрагмента Брауновского корпуса: «*The jury further said in term-end presentations 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 presentations_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 (знак \* означает, что следующая за ним буква прописная).

С точки зрения структуры лингвистических данных, различают форматы:

- с ключевыми словами;
- позиционный;
- гибридный.

### 3.2.1.1. XML-формат (формат с ключевыми словами)

Разметка с гибкой структурой предполагает запись грамматической информации в виде ключевых слов, соответствующих грамматическим категориям, с их значениями. Эти ключевые слова (категории) могут быть иерархически структурированы (вложены). В качестве универсального средства описания такой структуры сегодня чаще всего используется формализм XML. Многие системы разметки сохраняют результат своей работы именно в данном формате. Этот способ представления обеспечивает возможность явного документирования набора атрибутов и разделяет разметку структуры документа, его содержания и представления пользователю.

Приведем также в качестве примера морфологическую разметку фрагмента текста на русском языке («Звонили к вечерне. Торжественный гул колоколов...») в XML-формате на основе разметчика сервиса «Автоматическая обработка текста» (АОТ) (рис. 3.1).

```
<?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. 3.1. Пример морфологической разметки текста на русском языке в XML-формате

Файл разметки начинается со стандартной для XML-файлов строки, в которой указываются версия XML и кодировка текста. Сама разметка начинается с тега `<text>` и заканчивается тегом `</text>`<sup>4</sup>. Предложения описаны тегами `<s>`, а отдельные слова — тегами `<w>`.

После открывающего тега `<w>` идет исходная словоформа. Единицей морфологической разметки является слово (тег `<w>`). Исход-

<sup>4</sup> Здесь и далее названия тегов и атрибутов взяты из морфологической разметки системы АОТ ([www.aot.ru](http://www.aot.ru)).

ная форма, употребленная в тексте, записывается после этого тега. Морфологический разбор слова записан в элементе `<ana>`, у которого есть атрибуты:

- lemma — словарная форма в верхнем регистре;
- pos — часть речи (Г — глагол, ПРЕДЛ — предлог, С — существительное, П — прилагательное);
- gram — морфологические признаки. Грамматический код «gram» содержит в себе все возможные граммемы для данной части речи, обозначаемые как сокращения и перечисленные через запятую. В рамках этого грамматического кода возможна неоднозначность, когда, например, указывается сразу несколько возможных падежей.

Набор морфологических признаков для каждой части речи свой, причем значение признака неявно включает в себя название грамматической категории:

1	мн	Множественное число
2	нс	несовершенный вид
3	нп	непереходный
4	дст	действительный залог
5	приш	прошедшее время
6	жр	женский род
7	дт	дательный падеж
8	пр	предложный падеж
9	но	неодушевленное
10	кр	краткое прилагательное

Каждое слово может иметь одновременно несколько параллельных разборов, представленных в последовательности элементов `<ana>`. После разрешения неоднозначности (ручного или автоматизированного) в выходном представлении остается только один разбор.

Разметка на языке XML (с ключевыми словами) увеличивает объем файла разметки, но легко читается и воспринимается человеком без каких-либо вспомогательных инструментов и дает возможность свободно вводить дополнительную информацию. Например, в формате разметки, предложенном системой АОТ, на

очередном этапе появился атрибут *stress* (ударение). К недостаткам этого формата следует отнести невозможность сразу, без конвертирования, использовать его в существующих корпусных менеджерах.

### 3.2.1.2. Позиционный формат кодирования данных разметки

Позиционный формат разметки задает жесткую единообразную структуру полей данных, не допускающую никаких отклонений и вариантов.

Формат разметки, показанный на рис. 3.1, можно назвать линейным. Существует также вертикальный формат, где каждому токену соответствует одна строка, в которой следом за токеном записывается его морфосинтаксический признак или группа признаков (рис. 3.2).

Семеняка	Npfsny--	Семеняка
не	Qs	не
играла	V-is-sfa-p---	играть
в	Sps-	в
эти	Pd-раа--	этот
игры	Ncfран--	игра
.	.	.

Рис. 3.2. Пример позиционной разметки в вертикальном формате

Данная разметка представляет собой текстовый файл, в котором данные представлены в три столбца. В первом столбце приведены исходные словоформы, во втором — их грамматический код, в третьем — леммы (первоначальные (базовые) формы слова), соответствующие словоформе в данной строке. Фактически это таблица, ячейки которой разделены символом табуляции.

Наибольший интерес в этом формате представляет собой как раз грамматический код — именно он и является позиционным. Позиционность заключается в том, что для каждой части речи длина кода (количество символов в коде) строго регламентирована, а каждая грамматическая характеристика кодируется определенным символом в предписанной для нее позиции в этом коде (считая слева направо). Становится понятно, что в этом формате не остается места для неоднозначности, так как и грамматический

код, и грамматическая характеристика внутри кода могут быть представлены только один раз. Это значит, что перед тем как сформировать грамматический код и поместить его в файл, морфологический анализатор обязан сделать однозначный выбор в пользу какого-то одного варианта анализа, то есть снять грамматическую омонимию.

В позиционном формате для каждой грамматической характеристики отведен специальный участок (позиция) в поле грамматических признаков. В табл. 3.1 приведен пример позиционной морфологической разметки Чешского национального корпуса, разработанной в Институте формальной и прикладной лингвистики Карлова университета в Праге [Hajič, 2004]. Каждая словоформа описывается позиционным тегом (кодом) фиксированной длины, каждый тег содержит 15 ASCII-символов (в разметке Чешского национального корпуса 16 символов). Каждая позиция кодирует одну грамматическую категорию, некоторые позиции для определенных частей речи являются пустыми, некоторые зарезервированы — в обоих случаях они маркируются дефисом. Первая позиция всегда кодирует часть речи. Отличительной особенностью этого множества тегов является подробная детализация частей речи (позиция 2). Например, у числительных имеется 16 лексико-семантических типов, у местоимений — 21 тип и т. д.

Таблица 3.1. Пример позиционной морфологической разметки Чешского национального корпуса

Словоформа	Лемма	Морфологический тег
staví	stavit	VB-S---3P-AA---P
z	z	RR—2-----

Пример описывает разметку словосочетания *staví z* (строит из), где

- в первой строке:
  - staví (строит) — входная словоформа (токен)
  - stavit (строить) — лемма
  - VB-S---3P-AA---P — 16-значная морфологическая характеристика словоформы «staví»:
  - позиция 1 — часть речи: глагол (V)

позиция 2 — детализация части речи: глагол в настоящем или будущем времени (В)  
позиция 3 — род: не определено (-)  
позиция 4 — число: единственное (S)  
позиция 5 — падеж: не определено (-)  
позиция 6 — «притяжательный» род: не определено (-)  
позиция 7 — «притяжательное» число: не определено (-)  
позиция 8 — лицо: 3-е лицо (3)  
позиция 9 — время: настоящее время (P)  
позиция 10 — степень: не определено (-)  
позиция 11 — отрицание: утверждение (A)  
позиция 12 — залог: действительный (A)  
позиция 13 — резерв: не определено (-)  
позиция 14 — резерв: не определено (-)  
позиция 15 — вариативность (стиль): не определено (-)  
позиция 16 — вид: совершенный (P)

- во второй строке:

z (из) — входная словоформа (токен)

z (из) — лемма

RR--2----- — 16-значная морфологическая характеристика словоформы «z»:

позиция 1 — часть речи: предлог (R)  
позиция 2 — детализация части речи: обычный предлог (R)  
позиция 3 — род: не определено (-)  
позиция 4 — число: не определено (-)  
позиция 5 — падеж (которым управляет предлог): родительный (2)  
позиция 6 — «притяжательный» род: не определено (-)  
позиция 7 — «притяжательное» число: не определено (-)  
позиция 8 — лицо: не определено (-)  
позиция 9 — время: не определено (-)  
позиция 10 — степень: не определено (-)  
позиция 11 — не определено (-)  
позиция 12 — залог: не определено (-)  
позиция 13 — резерв: не определено (-)  
позиция 14 — резерв: не определено (-)  
позиция 15 — вариативность (стиль): не определено (-)  
позиция 16 — вид: не определено (-)

### 3.2.1.3. Гибридный формат кодирования данных разметки

Существует также гибридная (позиционно-атрибутивная, позиционно-фасетная) разметка, где морфосинтаксические признаки записываются в позиционном формате переменной длины. В табл. 3 приведены данные, сформированные морфологическим анализатором для чешского языка — *ajka* [Sedláček, Smrž, 2001], разработанным на факультете информатики Университета имени Масарика в Брно (Чехия)<sup>5</sup>. Это атрибутивная (фасетная) кодировка с начальным двухбуквенным кодом (ключевым словом) для частей речи и однобуквенным кодом для атрибутов (фасетов, категорий), следом за которыми идут их значения. Коды (теги) имеют разную длину в зависимости от числа атрибутов-категорий, уместных для той или другой части речи. Тот же самый пример в позиционно-атрибутивной (позиционно-фасетной) разметке будет выглядеть следующим образом (табл. 3.2):

- в первой строке:
 

staví (строит) — входная словоформа (токен)  
   stavit (строить) — лемма  
   k5eAaImIp3nP — морфологическая характеристика словоформы «staví», где:  
     k5 — глагол  
     e — отрицание: утверждение (A)  
     a — вид: несовершенный (I)  
     m — тип: индикатив настоящего времени (I)  
     p — лицо: третье (3)  
     n — число: множественное (S)
- и во второй строке:
 

z (из) — входная словоформа (токен)  
   z (из) — лемма  
   k7c2 — морфологическая характеристика словоформы «z», где:  
     k7 — предлог  
     с — падеж: родительный (2).

Стандарты для морфологической разметки еще не сложились окончательно. Среди конкурирующих друг с другом стандартов наиболее значимыми являются EAGLES, TEI и XCES (XML Corpus Encoding Standard).

---

<sup>5</sup> Более поздняя версия программы носит название «Majka» [Šmerk, 2009; Šmerk, 2010].

Таблица 3.2. Пример позиционно-фасетной морфологической разметки

Словоформа	Лемма	Морфологический тег
staví	stavit	k5eAaImIp3nS
z	z	k7c2

Правила EAGLES задают общие принципы создания и документирования корпусов и их морфосинтаксической разметки, а также ряд конкретных решений для разметки определенных случаев. EAGLES также допускает обе возможности морфологической разметки: или каждый признак представлен отдельным атрибутом (POS = 'NN', number = 'sing' и т. д.), или можно использовать позиционную кодировку, в которой закодированы значения категорий, 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 полного набора тегов для создания корпуса правила EAGLES не содержат.

Существующие корпусы, лингвистическая разметка которых основана на XML, используют самые разные системы кодирования. Например, BNC использует формат CDIF, основанный на TEI; Croatian National Corpus использует формат XCES; ICE (International Corpus of English), Czech National Corpus и Hungarian National Corpus применяют наиболее широко используемый стандарт TEI. Однако разные корпусы берут из TEI разные подмножества меток.

Приведем два примера разметки в TEI на морфологическом (1) и словообразовательном (2) уровнях:

- ```
(1) <w lemma='i' feats='pp1'>I</w>
    <w>
      <w lemma='do' feats='vvd'>did</w>
      <m type='negation'>n't</m>
    </w>
    <w lemma='do' feats='vv0'>do</w>
    <w lemma='it' feats='pp3'>it</w>
```
- 
- ```
(2) <w type='adjective'> comfortable
    <m type='prefix' baseform='con'>com</m>
    <m type='root'>fort</m>
    <m type='suffix'>able</m>
  </w>
```

Наиболее разработанным стандартом для лингвистической разметки текстов является стандарт XCES [Ide, Romary, 2002], который послужил основой международного стандарта в рамках проекта ISO TC37/SC4. XCES задает абстрактную *метамодель*, 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: *as well as* в английском или глаголы с отделяемыми приставками, например *zunehmen* в немецком, а также проводить декомпозицию одного слова в пределах разметки — описывать *zum* как *zu dem* в немецком.

### 3.2.2. Синтаксическая разметка

Синтаксическая разметка является результатом парсинга, выполняемого на основе данных морфологического анализа. Этот вид разметки зависит от принятой формальной синтаксической модели и описывает синтаксические связи между лексическими единицами и/или различные синтаксические конструкции (придаточное предложение, именное сказуемое и т. п.).

Автоматический синтаксический анализ (парсинг) — процесс сопоставления линейной последовательности лексических единиц текста с формальной грамматикой языка. В отличие от морфологии, способы представления синтаксической структуры и синтаксических отношений здесь не столь унифицированы. Наблюдается разнообразие синтаксических теорий и формализмов:

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

Как правило, результатом синтаксического анализа является формализованная синтаксическая структура предложения либо

в виде дерева зависимостей, либо в виде дерева составляющих, либо в виде некоторой комбинации первого и второго способов. Существуют и другие способы представления синтаксической структуры, например разбор предложения по членам в корпусе ХАНКО.

**Грамматика непосредственно составляющих** оперирует единицами, которые называются непосредственно составляющими (НС). НС — это, как правило, два элемента, из которых непосредственно образована единица более высокого порядка. Всем предложениям присуща линейная структура. Каждая составляющая более низкого уровня является частью составляющей более высокого уровня. Анализ по непосредственным составляющим может вестись «снизу» и «сверху». В последнем случае он соответствует традиционной процедуре разбора предложения: делению на подлежащее и сказуемое и описанию (разложению) группы каждого из них в терминах слов, словосочетаний и подчиненных предложений.

**Грамматика зависимостей** — формальная модель, представляющая структуру предложения в виде иерархии компонентов, между которыми установлено отношение зависимости (подчинения). Дерево зависимостей — иерархическая структура предложения, в которой все связи в предложении рассматриваются как подчинительные при наличии корня дерева — единственной вершины, которая никому не подчинена. В качестве такой вершины обычно признается сказуемое.

Разметка дерева зависимостей включает именование разных типов синтаксических отношений (определительное, обстоятельственное и т. п.); разметка дерева непосредственных составляющих подразумевает именование возникающих в процессе деривации сложных синтаксических объектов (именная группа, предложная группа и т. п.).

На рис. 3.3 изображено дерево зависимостей для английского предложения *John hit the ball*, где вершиной дерева является глагол *hit*. Горизонтальная (линейная) запись предложения может быть представлена в виде вертикально расположенного дерева зависимостей. Линейное представление деревьев см. ниже в виде XML-записи Глубоко аннотированного корпуса русского языка.

В русской компьютерной лингвистике распространена именно грамматика зависимостей. Разметка дерева зависимостей включает именование разных типов синтаксических отношений (определительное отношение, обстоятельственное отношение и т. п.).

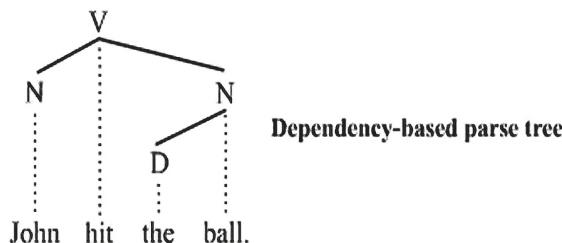


Рис. 3.3. Дерево зависимостей для английского предложения *John hit the ball* (пример из Википедии)

Формат представления структур зависимостей покажем на примере синтаксического корпуса проекта ЭТАП-3, который в настоящее время одновременно является синтаксическим корпусом НКРЯ (название корпуса, принятое в НКРЯ, ГАК — Глубоко аннотированный корпус русского языка). Глубоко аннотированный корпус организован следующим образом. Каждый входящий в него текст представляет собой отдельный файл в XML-формате, который содержит морфологическую информацию обо всех словоформах данного текста (то есть имя лексемы, соответствующей данной словоформе, и набор ее грамматических характеристик), а также синтаксическую структуру каждого предложения в виде дерева зависимостей [Апресян, Богуславский, Иомдин и др., 2005]. Такое представление о синтаксической структуре предложения восходит к лингвистической модели «Смысл  $\Leftrightarrow$  Текст» И. А. Мельчука и А. К. Жолковского. Перечень синтаксических отношений, используемых в ГАК, а также целый ряд конкретных лингвистических решений, связанных с представлением синтаксической структуры предложения, были выработаны в лаборатории компьютерной лингвистики Института проблем передачи информации РАН. В отличие от морфологически размеченного фрагмента НКРЯ, ГАК целиком состоит из структур со снятой морфологической и синтаксической омонимией.

Рассмотрим, как в анализаторе проекта ЭТАП-3 представлена структура следующего предложения:

*Как правило, для истинных ученых университет становится вторым домом, любовью на всю жизнь.*

На рис. 3.4 структура этого предложения показана в графической форме. Вершиной дерева является сказуемое *становится*.

Здесь дерево расположено горизонтально, оно как бы лежит на боку (ср. с вертикальной формой на рис. 3.3).



Рис. 3.4. Структура синтаксически размеченного предложения в виде дерева

Как видно на рисунке, в узлах дерева зависимостей стоят слова предложения, представленные леммами и цепочками грамматических характеристик (справа в квадратных скобках), а ветви помечены именами синтаксических отношений (в овалах). Морфологические теги и синтаксические связи (отношения) описаны в документации НКРЯ (<http://ruscorpora.ru/instruction-syntax.html>, раздел «Синтаксис»).

Ниже эта же структура предложения представлена в виде линейной XML-записи. Для наглядности повторим предложение еще раз.

*Как правило, для истинных ученых университет становится вторым домом, любовью на всю жизнь.*

```

<S ID="5">
  <W DOM="7" FEAT="CONJ" ID="1" LEMMA="КАК" LINK="вводн">
    Как</W>
    <W DOM=>1 FEAT=>S ЕД СРЕД ИМ НЕОД ID=>2 LEMMA=>ПРАВИЛО>
      LINK=>сравн-союзн>правило</W>,
      <W DOM="7" FEAT="PR" ID="3" LEMMA="для" LINK="обст">для</W>
      <W DOM="5" FEAT="А МН РОД" ID="4" LEMMA="ИСТИННЫЙ" LINK="опред">истинных</W>
      <W DOM="3" FEAT="S МН МУЖ РОД ОД" ID="5" LEMMA="УЧЕНЫЙ" LINK="предл">ученых</W>
      <W DOM=>7 FEAT=>V НЕСОВ ИЗЪЯВ НЕПРОШ ЕД 3-Л ID=>6 LEMMA=>УНИВЕРСИТЕТ>
        LINK=>предик>университет</W>
        <W DOM=>_root FEAT=>V НЕСОВ ИЗЪЯВ НЕПРОШ ЕД 3-Л ID=>7 LEMMA=>СТАНОВИТЬСЯ>становится</W>
  
```

```

<W DOM=>9» FEAT=>A ЕД МУЖ ТВОР» ID=>8» LEMMA=>ВТОРОЙ»
LINK=>определ>>вторым</W>
    <W DOM=>7» FEAT=>S ЕД МУЖ ТВОР НЕОД» ID=>9» LEMMA=>ДОМ»
LINK=>присвязь>>домом</W>,
        <W DOM=>9» FEAT=>S ЕД ЖЕН ТВОР НЕОД» ID=>10» LEMMA=>
любовь» LINK=>сочин»>любовью</W>
            <W DOM=>10» FEAT=>PR ID=>11» LEMMA=>ХА» LINK=>атриб»
>на</W>
                <W DOM=>13» FEAT=>A ЕД ЖЕН ВИН» ID=>12» LEMMA=>ВЕСЬ»
LINK=>определ»>всю</W>
                    <W DOM=>11» FEAT=>S ЕД ЖЕН ВИН НЕОД» ID=>13» LEMMA=>ЖИЗНЬ»
LINK=>предл»>жизнь</W>.
                </S>

```

Каждому слову в предложении (тег `<W>`) приписывается идентификатор (ID — порядковый номер), лемма (LEMMA), набор граммем (FEAT), идентификатор слова-хозяина (DOM) (для корня дерева (глагольная форма *становится*) DOM=ROOT) и тип связи (LINK).

Рекомендации TEI содержат свои теги для XML-записи синтаксических отношений, а именно:

- тег `<c1>` — клауза для кодирования сложносочиненных и сложноподчиненных предложений, у него есть два атрибута: *type*, задающий синтаксические признаки клаузы, 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 структуры в терминах зависимостей предусмотрены специальные теги, например `<dep>`, который имеет атрибуты *function* и *target*, последний ссылается на идентификатор зависимого слова в предложении.

Приведем пример морфосинтаксической разметки в TEI фрагмента предложения *Nineteen fifty-four, when I was eighteen years old*:

```

<p>
<c1 type='finite declarative' function='independent'>
<phr type='NP' function='subject'>Nineteen fifty-four,
<c1 type='finite relative declarative'
function='appositive'>when
<phr type='NP' function='subject'>I</phr>
<phr type='VP' function='predicate'>was eighteen
years old</phr>
</c1>,
</phr>...

```

### **3.2.3. Семантическая разметка**

Семантические теги могут обозначать 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. п.

Один из вариантов семантической разметки предлагает Национальный корпус русского языка (НКРЯ). В этом корпусе каждой словоформе приписываются пометы трех типов:

- разряд (имя собственное, возвратное местоимение и т. п.);
- лексико-семантические характеристики (тематический класс лексемы, признаки каузативности, оценки и т. п.);
- деривационные характеристики (диминутив, отадъективное наречие и т. п.).

Собственно лексико-семантические теги 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:

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

Другой вариант семантических тегов для русской семантической разметки представлен на сайте Центра компьютерных и корпусных языковых исследований Ланкастерского университета (University Centre for Computer Corpus Research on Language) (<http://ucrel.lancs.ac.uk/usas/>) (табл. 3.3). Начало таблицы выглядит следующим образом:

Таблица 3.3. Русская семантическая разметка

A	Общие понятия
A1	Общие понятия
A1.1	Обычные действия / Изготовление ч.-л.
A1.2	Повреждение и разрушение
A1.2	Пригодность
A1.3	Осторожность
A1.4	Шансы, везение
A1.5	Употребление, применение, использование
A1.5.1	Использование
A1.5.2	Польза
A1.6	Материальное/нематериальное
...	.....
B	Тело человека
B1	Анатомия и физиология

Окончание табл. 3.3

B2	Здоровье и болезнь
...	.....
C	Искусства и ремесла
...	.....

Вся система семантических категорий представлена на сайте Центра и подробно описана в [Arcer Wilson, Rayson, 2002].

Семантические теги показывают семантические поля, которые объединяют значения слов, характеризующих один и тот же концепт с разной степенью обобщения. Эти группы включают не только синонимы и антонимы, но и гиперонимы и гипонимы.

Программа USAS представляет собой рабочую среду для проведения автоматического семантического анализа текста. Она была разработана в Ланкастерском университете в 1990 г. и используется для осуществления серии исследовательских проектов. В частности, в ней есть семантический теггер для английского языка, который автоматически размечает текст с точки зрения присвоения семантических тегов при горизонтальном просмотре (рис. 3.5) и присвоения морфологических и семантических тегов при вертикальном просмотре (рис. 3.6). Например, разметка предложения *Written and spoken data produced by learners has always been a key resource for the study of second language acquisition (SLA)* будет выглядеть следующим образом:

```
Written_Q1.2 and_Z5 spoken_Q2.1 data_X2.2/X2.4
produced_A2.2 by_Z5
learners_P1/S2mf has_Z5 always_N6+++ been_A3+ a_Z5
key_A11.1+ resource_A9+
for_Z5 the_Z5 study_P1 of_Z5 second_N4 language_Q3
acquisition_A9+ (_PUNC
SLA_Z99 )_PUNC ._PUNC
```

Рис. 3.5. Семантические теги при горизонтальном просмотре

Существуют и другие типы разметки, в частности:

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

- просодическая разметка. В просодических корпусах применяются теги, обозначающие ударение и интонацию. В корпусах устной разговорной речи просодическая разметка часто сопровождается *дискурсной* разметкой, которая служит для обозначения пауз, повторов, оговорок и т. п.

0000001	002	-----	-----
0000003	010	VVN	Written Q1.2 I2.1
0000003	020	CC	and Z5
0000003	030	JJ@	spoken Q2.1
0000003	040	NN	data X2.2/X2.4 Q1.1
0000003	050	VVN	produced A2.2 A1.1.1 A10+ K4 K3 Q4.3 F4
0000003	060	II	by Z5
0000003	070	NN2	learners P1/S2mf
0000003	080	VHZ	has Z5 A9+ A2.2 S4
0000003	090	RR	always N6+++
0000003	100	VBN	been A3+ Z5
0000003	110	AT1	a Z5
0000003	120	JJ	key A11.1+
0000003	130	NN1	resource A9+ W3
0000003	140	IF	for Z5
0000003	150	AT	the Z5
0000003	160	NN1	study P1 X2.4 H2 Q1.2 C1
0000003	170	IO	of Z5
0000003	180	MD	second N4
0000003	190	NN1	language Q3 Y2
0000003	200	NN1	acquisition A9+
0000003	201	(	(
0000003	210	NP1	SLA Z99
0000003	211	)	)
0000003	212	.	.

Рис. 3.6. Морфологические и семантические теги при вертикальном просмотре

### 3.3. Экстралингвистическая разметка

Экстралингвистическая разметка (метаразметка) включает в себя внешнюю, «интеллектуальную» разметку (библиографические характеристики, типологические, тематические, социологические характеристики), которую могут дополнять данные технологической разметки (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 явно или неявно основан на рекомендациях проекта TEI. Выделяют два класса факторов, влияющих на язык текстов:

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

Дж. Синклер [Sinclair, 1996] 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) [Sinclair, 1996].

В НКРЯ, например, используется следующий набор метаданных.

Первый блок:

- *автор текста*: имя, пол, дата рождения (или примерный возраст);
- *название текста*;
- *время и место создания текста* (может указываться точно или приблизительно);
- *объем текста*: для художественных произведений принято, 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:

- жанр текста: нежанровая проза, автобиографическая проза, детектив, детская литература, историческая проза, криминальная литература, приключения, фантастика, юмор и сатирик;
- тип текста: автобиографическая проза, анекдот, ассоциативная проза, боевик, детектив, очерк, литературное письмо, повесть, притча, пьеса, рассказ, роман, сказка, триллер, эпopeя, эссе и др.;
- хронотоп текста: приблизительное указание на место и время описываемых в тексте событий [Национальный корпус русского языка].

При указании на хронотоп текста в НКРЯ предлагаются следующие опции: Древний Восток; Россия XVII в.; Россия XIX в.; 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 тег «хронотоп не определен».

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

- «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);
- аудитория — возраст;
- аудитория — уровень образования;
- аудитория — размер.

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

## Глава 4. Типология корпусов

### 4.1. Классификация корпусов по различным основаниям

Несмотря на разнообразие корпусов, прежде всего можно выделить два основных основания их деления на классы:

- 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 учёного и т. д.;
- 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.

Можно выделить большое число разных типов корпусов в зависимости от исследовательских и прикладных задач, для решения которых они создаются, и различных оснований для классификации. В зависимости от поставленных целей и классифицирующих признаков можно выделить различные типы корпусов (табл. 4.1). 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 [Митрофанова, Захаров, 2009]. В этом корпусе методология корпусной лингвистики применена к ней самой.

Корпусы текстов могут классифицироваться по жанрам и подразделяться на литературные, фольклорные, драматургические,

публицистические и др. Примером публицистического корпуса может служить *Компьютерный корпус текстов русских газет конца XX в.* (<http://www.philol.msu.ru/~lex/corpus/>).

Таблица 4.1. Классификация корпусов

Признак	Типы корпусов
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»

В современных информационных технологиях понятие жанра оказывается размытым. Уже ряд лет широко обсуждается проблема

идентификации и определения веб-жанров [Mehler, Sharoff, Santini, 2010], имеющая непосредственное отношение и корпусной лингвистике.

По назначению выделяют исследовательские и иллюстративные корпусы. Исследовательские корпусы создаются с целью изучения различных аспектов функционирования языка. Этот тип корпусов ориентирован на широкий класс лингвистических задач. Неспецифицированность задачи требует корпусы достаточно большого объема. Как правило, такие корпусы текстов содержат от нескольких сотен миллионов до нескольких десятков миллиардов 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 хороших лингвистических примеров, подтверждающих те или иные языковые (речевые, текстовые) факты, обнаруженные ранее иными лингвистическими приемами [Баранов, 2007].

Типичный пример иллюстративного корпуса представлен в «Путеводителе по дискурсивным словам русского языка» [Баранов, Плунгян, Рашилина, 1993], в котором семантический анализ частиц и выделенные значения сопровождаются обширным текстовым материалом, что позволяет читателю проверить семантические интерпретации, предложенные авторами.

В системе Sketch Engine разработаны подсистемы GDEX (Good Dictionary Examples) и SkELL (Sketch Engine for Language Learning). GDEX — это автоматизированная система оценки предложений с точки зрения их пригодности для использования в качестве словарных примеров, которая базируется на корпусных данных. Вся современная лексикография построена на корпусах, однако поиск в современных корпусах дает тысячи и десятки тысяч контекстов. Их просмотр и отбор хороших примеров занимают огромное время. GDEX эффективно исключает «плохие» примеры и предлагает лексикографу набор предложений, которые с большой вероятностью содержат хорошие предложения в качестве словарных примеров. В системе используется довольно сложная многоступенчатая эвристика отбора хороших примеров. Конфигурация GDEX настраивается с учетом конкретного языка, а пользовательские конфигурации могут быть разработаны с учетом конкретной цели.

SkELL представляет собой веб-инструмент, основанный на корпусе, который позволяет изучающим язык и преподавателям находить аутентичные контексты, устойчивые сочетания и списки семантически связанных слов для конкретных целевых слов. SkELL базируется на инструментах корпусного менеджера Sketch Engine и модуле GDEX. Предложения отбираются из специального текстового корпуса, созданного на базе веба, очищенного от спама и содержащего только высококачественные тексты новостей, научных статей из Википедии и художественной литературы. Существуют версии SkELL для английского, русского, немецкого, итальянского, чешского и эстонского языков.

Критерий *динамичность* подразделяет корпусы на динамические и статические. Первоначально корпусы текстов создавались как статические образования, отражающие определенное временное состояние языковой системы. Статические корпусы содержат тексты какого-либо временного 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 временной шкале, например изменения значений слов, частоты использования тех или иных синтаксических конструкций и т. п. Для отражения процессуального аспекта проблемной области разрабатываются технологии построения и эксплуатации динамических корпусов текстов [Баранов, 2007]. Динамические корпусы называют также мониторными или мониторинговыми. Их цель — 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 неанно-

тированные, тегированные и нетегированные. В размеченном корпусе словам или предложениям присваиваются теги в соответствии с *характером разметки*: морфологические, синтаксические, семантические, просодические и др.

Важным критерием для пользователей является *доступность* корпуса. Свободно доступные корпусы позволяют в любое время в режиме онлайн 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 онлайн 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) [Богданова, 2010].

Рассмотрим более подробно еще три типа корпусов, которые заслуживают особого внимания. Критериями для выделения этих разновидностей корпусов являются, соответственно, *параллельность*, *тип языковых данных и назначение*.

## 4.2. Особенности корпусов отдельных типов

### 4.2.1. Параллельные корпусы

По критерию *параллельность* корпусы делятся на два основных типа:

- собственно параллельные корпусы (*translation corpora*) — корпусы, представляющие собой множество текстов-оригиналов, написанных на каком-либо исходном языке, и текстов — переводов этих исходных текстов на один или несколько других языков;
- сопоставимые корпусы (*comparable corpora*) — корпусы, объединяющие тексты по какому-либо признаку (одна и та же тематическая область, диалекты какого-либо языка, региональные варианты, разновидности английского или иного языка, такие как английский как родной и английский как иностранный, и т. д.).

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, диалектологии, переводоведения и т. п.).

При подготовке параллельных корпусов первого типа и разработке программ для их обработки возникает проблема выравнивания (*alignment*) — установления соответствий между фрагментами текста оригинала и текста перевода. Для решения этой задачи используются различные методы автоматического выравнивания текстов: по предложениям (табл. 4.2), 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, шесть возможных соответствий между предложениями обоих текстов:

- одно исходное предложение переводится одним предложением;
- два исходных предложения переводятся одним предложением;
- одно исходное предложение переводится двумя предложениями;

- два исходных предложения переводятся двумя предложениями, но внутренние границы этих предложений в тексте оригинала и в тексте перевода не совпадают;
- предложение исходного текста не переводится;
- предложение в тексте перевода не имеет эквивалента в тексте оригинала [Зубов, Зубова, 2004].

Таблица 4.2. Пример автоматического выравнивания по предложениям текста романа Иэна Макьюэна «Искупление» (Atonement) и его перевода И. Дорониной

1	THE PLAY — for which Briony had designed the posters, programs and tickets, constructed the sales booth out of a folding screen tipped on its side, and lined the collection box in red crepe paper — was written by her in a two-day tempest of composition, causing her to miss a breakfast and a lunch.	Пьеса, для которой Брайони рисовала афиши, делала программки и билеты, сооружала из ширмы кассовую будку и обклеивала коробку для денежных сборов гофрированной красной бумагой, была написана ею за два дня в порыве вдохновения, заставлявшего ее забывать даже о еде.
2	When the preparations were complete, she had nothing to do but contemplate her finished draft and wait for the appearance of her cousins from the distant north	Когда приготовления закончились, ей не оставалось ничего, кроме как созерцать свое творение и ждать появления кузенов и кузины, которые должны были прибыть с далекого севера

Еще большую сложность представляет собой выравнивание на уровне клауз и слов.

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, в диалоговом режиме или с постредактированием результатов автоматического выравнивания. В качестве примеров программ выравнивания можно назвать Hunalign (имеется графический интерфейс Euclid, разработанный в Высшей школе экономики), Abbyy Aligner, Trados, Winalign, Wordfast tools, Giza++ и др.

Параллельные корпусы могут быть одноязычными (если сопоставляются диалекты, варианты одного языка, язык носителей и изучающих данный язык), двухязычными и многоязычными.

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;
- устранивать полисемию лексических единиц путем компьютерного анализа контекста многозначного слова, причем контекст по длине может превышать предложение;
- переводить терминологические и фразеологические единицы текста;
- формировать семантические поля и терминологические системы;
- изучать универсалии перевода;
- 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 [Зубов, Зубова, 2004].

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 [Беляева, 2004]. 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 при определении адекват-

ности 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 [Митрофанова, Грачкова, Шиморина, 2010; Гарабик, Захаров, 2006].

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, русский и белорусский, сербский и хорватский. Системы переводческой памяти могут быть творчески использованы для большей автоматизации переводческого процесса, не зависящей от конкретных языков.

Параллельные корпусы часто создаются на основе текстов, используемых в многоязычных сообществах (Организация Объединенных Наций, Европейский союз) и в официально двуязычных странах (Канада, Финляндия, Индия и др.).

#### **4.2.2. Корпусы устной речи**

*По типу языковых данных* корпусы делятся на письменные, устные (речевые или звуковые) 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 разных формах в определенный период времени (НКРЯ, BNC и др.).

Составители корпуса не всегда представляют 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 устной речи не была так тщательно исследована в компьютерной лингвистике и корпусных исследованиях, как некоторые другие сферы лингвистики, поскольку создание презентативного корпуса устной речи было сложной задачей. Однако развитие диалоговых интернет-сервисов, необходимость в создании моделей вежливости, смены ролей и других явлений [Finegan, 2004] потребовали обратить особое внимание на этот тип корпусов.

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 университете колледже. Объем корпуса — 1 млн словоупотреблений. Текстами устной речи были записи радиопередач, заседаний официальных структур, а также неформальных бесед.

Машинный вариант корпуса объемом 500 тыс. словоупотреблений 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» [Svartvik, Quirk, 1980]. Эта книга была очень полезна в то время, когда компьютерные корпусы не были широко распространены и было трудно обращаться со сложной транскрипцией устной речи. 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 изучался в сопоставительных исследованиях на базе корпусов Лондон-Лунд и Ланкастер-Осло-Берген; в частности, изучались модальность, связи в сложных предложениях, отрицание.

Отсутствие баланса в доступности устного и письменного материала в машиночитаемом формате будет давать знать о себе 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, видеокассеты или цифровой записи. Затем ее нужно записать буквами, что является утомительной и дорогой работой, качество которой во многом зависит от качества имеющейся записи и уровня шума внешней среды в естественных условиях.

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

- выбор алгоритма для транскрибирования;
- учет индивидуальных особенностей произношения;
- учет всего устного текста или его фрагментов;
- учет диалектных вариантов произношения слов;
- учет ударений в словах;
- учет просодических признаков произносимых фраз;
- маркирование слов, которые при прослушивании не распознавались;
- маркирование паралингвистических явлений, сопутствующих речи (пауз, смеха, бормотания, кашля и т. п.), в записи для фонетического корпуса.

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 достаточно много достаточно представительных фонетических корпусов.

Так, в 1970-х годах 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 учитывались. 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 [Зубов, Зубова, 2004].

Один из членов команды, создававшей Британский национальный корпус, Л. Бернард (L. Burnard), утверждал, что стоимость отбора 10 млн слов из устных источников во время создания корпуса (1990-е годы) равнялась стоимости отбора 50 млн слов из письменных источников [Николаева, 2010]. Данные издержки связаны еще и со строго соблюдаемым в западном мире авторским правом, в связи с чем нельзя провести полноценного анализа устных текстов и опубликовать его результаты без получения согласия их автора, а это не всегда возможно по объективным причинам.

#### **4.2.3. Учебные корпусы текстов**

Традиционно учебные корпусы текстов — Learner corpora (LC) — используются в рамках теории овладения вторым языком и зачастую представляют собой аннотированные корпусы ошибок, которые допускают школьники или студенты в процессе его изучения под руководством преподавателя. Такие корпусы текстов созданы в Бельгии (The International Corpus of Learner English), в Финляндии (English as Lingua Franca in Academic Setting), в России (Корпус английских текстов петербургских школьников), во Франции (French learner writing corpus) и в других странах. Подобные корпусы текстов могут быть с успехом использованы для выявления и анализа наиболее распространенных ошибок в изучаемых языках. Большой объем лингвистического материала, содержащегося в таких корпусах текстов, и возможности современных информационных техно-

логий обработки текстов должны значительно облегчить и ускорить процесс создания новой учебной литературы и процесс отбора демонстрационного материала непосредственно для обучения в аудитории или классе.

Создание учебных корпусов преследует как педагогические, так и исследовательские цели. С одной стороны, на базе таких корпусов могут создаваться лингвистические тренажеры, тестовые и контрольные задания (с учетом самых распространенных ошибок), индивидуальные учебные подкорпусы студентов. С другой стороны, ошибки служат маркерами языковых изменений.

Первым и самым известным учебным корпусом является Международный учебный корпус английского языка (The International Corpus of Learner English — ICLE). Он включает в себя аргументативные эссе, написанные студентами 3–4-х курсов продвинутого языкового уровня. Основная цель ICLE — исследование языка межнациональной коммуникации студентов, изучающих английский язык.

Корпус английских текстов петербургских школьников, предназначенный в основном для исследования особенностей английских текстов, порождаемых школьниками, создавался на кафедре прикладной лингвистики РГПУ им. А. И. Герцена под руководством О. Н. Камшиловой [Камшилова, 2010; 2012]. Большинство известных учебных корпусов фиксируют определенный этап языковой компетенции. Новое направление в учебных корпусах — создание лонгитюдных корпусов, накопление текстов одного и того же автора (авторов) в течение некоторого времени, что позволяет представить процесс овладения языком в динамике [Камшилова, 2016; Камшилова, Захаров, 2017].

В настоящее время создаются и такие учебные корпусы текстов, которые имеют целью выявить наиболее распространенные типы ошибок в конструкциях, в первую очередь помеченные тегами *cause*, *contam* и *lex*, *phrase*, и определить причины ошибок в конструкциях, которые совершают студенты в учебных текстах (эссе, дипломных и курсовых работах и т. п.). Корпус русских учебных текстов (КРУТ) (объем — более 2,5 млн словоупотреблений) существует с 2013 г. как один из проектов Лингвистической лаборатории по корпусным технологиям на факультете филологии в НИУ ВШЭ под руководством Е. В. Рахилиной [Пужаева, 2015]. КРУТ — открытый бесплатный интернет-ресурс (<http://web-corpora.net/CoRST/search/>), содержащий

учебные тексты студентов НИУ ВШЭ и других вузов, носителей русского языка. Исследование проводится в рамках двух направлений современной лингвистики — грамматики конструкций и грамматики ошибок.

### **Вопросы и задания для самоконтроля**

1. Дайте определения терминов: *корпус, корпусная лингвистика, разметка, препрезентативность, лемма, метаданные*.
2. Какие три типа корпусов можно выделить по критерию прагматической ориентированности? В чем их отличие?
3. Какой корпус текстов был первым? Укажите его основные характеристики.
4. Какие стандарты действуют в корпусной лингвистике? Что подлежит стандартизации?
5. Какие типы лингвистической разметки существуют?
6. Что представляет собой экстравалингвистическая разметка?
7. Перечислите типы корпусов.
8. Каково основное назначение параллельных корпусов текстов?
9. В чем заключается сложность создания корпусов устной речи?
10. Что такое лонгитюдный корпус?

## Часть 2

# Создание корпусов

## Глава 5. Традиционная технология создания корпусов

### 5.1. Проектирование и технологический процесс создания корпусов

Проект любого корпуса должен предусматривать этапы создания и пути его дальнейшего развития. Понятие корпуса является продолжением традиционных картотек, с которыми всегда работали лингвисты. Значительную роль в становлении корпусного подхода сыграл Интернет, в процессе развития которого 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. 1.5.1), 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, иногда достаточно экзотических. 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, цитатами, прямой речью, иноязычными вкраплениями, единицами изменения и т. п.

Все эти вопросы должны быть поставлены на этапе проектирования. Решать же их (к сожалению, лишь некоторые из них) можно постепенно, в процессе создания и опытной эксплуатации корпуса. Для этого с самого начала эксплуатации следует предусмотреть обратную связь с пользователями и возможность внесения корректировок в процедуры обработки текстов.

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

1. Обеспечение поступления текстов в соответствии с перечнем источников.
2. Преобразование в машиночитаемый формат. Тексты в электронном виде для создания корпусов могут быть получены самыми разными способами — ручным вводом, сканированием, авторскими копиями, в дар или в обмен, через Интернет, оригинал-макетами, предоставляемыми составителям корпусов, и др.
3. Техническая подготовка текстов, которую выполняют вручном режиме и сопровождают библиографическим (металингвистическим) описанием текста.
4. Предварительная программная обработка текстов, например удаление или преобразование нетекстовых элементов (рисунков, таблиц), удаление из текста переносов, «жестких концов строк» (текстов из MS-DOS), обеспечение единого написания тире и т. д.
5. Токенизация, предполагающая проведение следующих операций: разделение входного текста на элементы (слова, разделители и т. д.), выделение и оформление нестандартных (нелексических) элементов, обработка специальных текстовых элементов (имен, написанных инициалами, иностранных лексем, записанных латиницей, названий рисунков, примечаний, страниц форзаца, зачеркваний, титульных листов, списков литературы и т. д.). Как правило, эти операции выполняются в автоматическом режиме. Обычно на этом же этапе осуществляется сегментирование текста на структурные составляющие (абзацы, предложения).
6. Разметка текста. Тексты и их компоненты получают дополнительную информацию (метаданные). Метаданные мож-

но поделить на три типа: экстралингвистические, относящиеся ко всему тексту; данные о структуре текста; лингвистические метаданные, описывающие элементы текста. Метаописание текстов корпуса включает прежде всего содержательные элементы данных, вносимых в ручном режиме (библиографические данные, признаки, характеризующие жанровые и стилевые особенности текста, сведения об авторе).

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 многоаспектный поиск и статистическую обработку.
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 можно предоставлять разные права и 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 корпусменеджера и т.д.

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

## 5.2. 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. Какова иерархическая структура корпуса?
4. Какие письменные текстовые источники должны быть представлены в корпусе текстов и в каком количестве?
5. Из какой исходной языковой области должны быть выбраны тексты, включаемые в состав корпуса?

В отечественной лингвистике ответы на эти вопросы отчасти 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. [Зубов, Зубова, 2004]. Схожие проблемы решались при создании «Частотного словаря русского языка» под руководством Л.Н. Засориной (см. предисловие к этому словарю) [Засорина, 1977].

Основными единицами корпуса текстов являются *словоупотребления* (или токен, англ. *tokens*, также их называют словами, *words*). Кроме того, корпусная лингвистика оперирует понятиями *словоформа* (*type*), *основная форма*, *лемма* (*lemma*), *оборот*, *устойчивое словосочетание* (*multiword expression*), *предложение* (*sentence*). Объем создаваемого корпуса текстов в принятых единицах зависит от целей создания. Кроме обычных слов, токены включают в свой состав знаки препинания, числа, марки и т. п. Объем корпуса может быть небольшим, если изучается частота употребления букв, буквосочетаний, звуков, звукосочетаний. Гораздо большим он должен быть при изучении лексики, морфологических явлений и синтаксических или стилистических особенностей текстов. Для изучения фразеологии требуются корпусы еще большего размера.

При создании корпуса проблемными являются 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 временных промежутков включать в национальный (общезыковой) корпус текстов (современные, написанные 20, 50 лет назад и более)?

3. Включать ли тексты только литературного языка или также другие типы источников? И что считать литературным языком?

Для того чтобы ответить на эти вопросы, разработчики корпуса обращаются за консультациями к специалистам по языкоznанию и лингвостатистике или используют 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 [Зубов, Зубова, 2004].

### 5.3. Основные процедуры обработки входных текстов

Основными процедурами обработки корпусных текстов следует считать процедуры, указанные выше под номерами 5 и 6 (разд. 5.1), — токенизацию и разметку.

В процессе создания корпуса естественно использование специальных процедур и программ. Например, **токенизация** (или графематический анализ), то есть разделение потока символов в текстах на естественном языке на отдельные значимые единицы (токены, словоформы, знаки препинания), является необходимым условием для дальнейшей обработки текстов. Если бы языки обладали совершенной и однозначной 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 токенизиро-

ваны. 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 как отдельный токен. Подобные примеры могут быть приведены и для других языков. Во многих языках приходится решать проблему дефиса, который может быть как конкатенатором, так и делимитатором (например русские слова *из-за*, *кто-то*, *кое-как*, *буквенно-цифровой*, *во-первых*, *плац-палатка* — и *девочка-пионерка*, *старик-извозчик*, *туристы-японцы*,  *завод-изготавитель*, *январь-февраль* и т. п.).

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

Одна из важных функций токенизации, которая в большинстве корпусов не реализована, — это выделение многословных лексических единиц. В их числе могут быть многословные служебные слова, например союз «потому что» или предлог «в связи с» (их не так много), аналитические формы, глаголы с отделяемыми приставками в немецком языке и прежде всего содержательные устойчивые многословные единицы: идиомы, составные термины, имена собственные, различные именованные сущности, которые в каждом языке представлены в большом количестве. В немецком языке имеется обратная проблема, которая должна решаться или на стадии токенизации, или в процессе поиска, — разделение сложных слов на составные компоненты.

Структурная разметка документа (выделение абзацев, предложений, слов) и токенизация обычно осуществляются автоматически.

За токенизацией следует процедура **разметки**, или правильнее сказать, разметок. Как правило, всегда выполняется *морфологическая разметка* (*tagging*), или частеречная разметка (*part-of-speech tagging*), за которой могут следовать синтаксическая, семантическая, поэтическая и др.

В задачу морфологической разметки входит **лемматизация**, то есть процедура образования первоначальной формы слова (леммы)

для словоформ текста, и приписывание единицам текста (токенам) **граммем** — значений грамматических и, возможно, других категорий. В качестве других категорий можно назвать, например, признак фамилии, признак многокомпонентной единицы, признак аналитической формы глагола.

Как правило, в большинстве языков слово встречается в нескольких формах. Например, английский глагол *walk* может быть представлен следующими формами: *walk*, *walked*, *walks*, *walking*. Тогда данной конкретной словоформе, помимо набора грамматических признаков, в качестве отдельного элемента метаописания будет приписана базовая форма *walk*, зафиксированная в словаре и называемая леммой слова.

*Синтаксическая разметка* (англ. *annotation*) базируется на синтаксическом анализе, часто называемом **парсингом**. Это процесс соединения линейной последовательности лексем (слов, токенов) языка с его формальной грамматикой. Результатом обычно является *дерево зависимостей* (синтаксическое дерево) или *дерево составляющих*, которые записываются в виде линейного файла со специальными тегами, описывающими древесную структуру (см. п. 3.2.2). Корпус с такой разметкой называется синтаксическим, или аннотированным, корпусом (англ. *treebank*).

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.

Еще большую сложность представляют другие виды лингвистической разметки.

Экстралингвистические метаданные (в полном объеме) являются результатом интеллектуальной обработки текстов и вводятся вручную. Их количество и качество во многом определяют возмож-

ности использования корпуса. При поиске метаданные используются или как поисковые признаки, или как приемы формирования подкорпуса, на котором будет выполняться запрос (что, в общем, одно и то же). Множество результатов поиска также может быть обработано и показано «через призму» метаданных, то есть в виде статистических характеристик выдачи, демонстрирующих распределение найденных документов по авторам, по сферам функционирования, по типам и тематике текста и по жанрам.

После выполнения вышеописанных и всех других процедур по обработке текстов осуществляется операция конвертирования размеченных текстов в базу данных корпусного менеджера.

#### **5.4. Как создать собственный корпус?**

Каждый хороший корпус — это результат большой многолетней работы специалистов высокой квалификации. Поэтому обычно удобнее пользоваться готовыми корпусами, тем более что рядовому пользователю вряд ли по силам создать что-то большое и мощное.

Следует различать два типа корпусных менеджеров: предоставляющих пользователю возможность загрузить в корпус свои данные и не дающих такой возможности. Большинство современных корпусов, а фактически корпусных менеджеров, относится ко второму типу. Это большие закрытые собрания корпусных данных, к которым возможен только доступ. Сегодня в Сети можно найти огромное количество самых разных корпусов.

Но иногда возникает необходимость в каком-то особом корпусе. Например, требуется создать корпус какого-то редкого языка. Или нужно провести машинное обучение с помощью данных, которые отсутствуют в стандартном корпусе. Например, для создания робота, управляющего трактором, понадобился корпус высказываний трактористов во время работы, чтобы робот правильно взаимодействовал с коллегами на соседних машинах (пример О. В. Митрениной). Во всех этих случаях нужно создавать корпус самостоятельно. Для этого существуют различные пути, один из которых — создание собственного корпусного менеджера. Но этот путь малоперспективен. Более целесообразно воспользоваться готовыми программно-лингвистическими средствами. Прежде всего нужно четко сформулировать задачу: какой корпус предстоит создать? При этом нужно понимать, что процесс создания кор-

пуса — это целый ряд процедур: токенизация, фильтрация, устранение дублей, морфологическая разметка, синтаксическая разметка, загрузка корпусных данных и др., каждая из которых нуждается в соответствующих программных средствах. Поэтому правильнее всего выбрать систему, где эти средства интегрированы в единый комплекс. Естественно, этот выбор требует определенных знаний, понимания цели и готовности идти на определенные компромиссы, то есть подстраиваться под выбранный инструмент. Бесплатных систем, позволяющих создавать свои корпусы (впрочем, и платных тоже), не так много. Среди них можно назвать корпусные менеджеры WordSmith Tools, CQP, Xaira, DDC, MonoConc, ParaConc и ряд других. Одни корпусные менеджеры, как правило более мощные, реализованы на архитектуре «клиент — сервер», другие могут быть установлены на локальный компьютер. В первом случае требуется установка и настройка сервера, установка и настройка программного обеспечения корпусного менеджера, поддержка сервера в течение всего срока реализации проекта. Все это, как правило, не под силу обычному лингвисту.

Для создания небольших корпусов (1–10 млн словоупотреблений) можно порекомендовать корпусный менеджер AntConc, который работает на локальном компьютере. В этой функционально мощной системе отсутствует модуль лемматизации и морфологической разметки. (Если можете без этого обойтись, то идите на компромисс и работайте.)

Для создания больших корпусов или для доступа к большой функциональности может быть использована система NoSketch Engine, разработанная в Университете Масарика в Брно, Чешская Республика (<https://nlp.fi.muni.cz/trac/noske>) [Rychlý, 2007]. Но в этом случае у пользователя должен быть свой сервер с установленной на нем операционной системой Linux и другими необходимыми библиотеками.

Если все это пользователю не по силам и при этом требуется провести исследование в сжатые сроки, мы рекомендуем воспользоваться услугами платной корпусной службы Sketch Engine, позволяющей создавать корпусы как на основе собственных текстов, так и по технологии WaC на основе текстов из веба.

Имеется также упрощенный сервис BootCaT (<https://bootcat.dip-intra.it/>), позволяющий создавать свои корпусы аналогичным способом (подробнее см. п. 6.2).

## Глава 6. Создание корпусов на базе веба

### 6.1. Поисковые системы Интернета как корпусы

Информационное наполнение Интернета (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 большую ценность для лингвистов, так как перевод текстов в машиночитаемый формат и создание корпусов требует больших временных и материальных затрат.

При использовании веб-пространства как корпуса для решения лингвистических задач роль корпусных менеджеров могут выполнять поисковые системы. Основным средством поиска информации в Сети являются глобальные информационные поисковые системы верbalного типа (поисковые машины — *search engines*), индексирующие все интернет-пространство и обеспечивающие поиск по тексту. Индексы (инвертированные файлы) поисковых систем — это, по сути, не что иное, как виртуальные конкордансы к текстам. Результаты поиска в информационных поисковых системах в виде кратких описаний документов, как правило, содержат контексты, в которых искомые слова встретились в найденных документах. При этом полезно понимать, как строятся эти индексы вербальных систем, что собой представляют языки запросов поисковых систем, и, соответственно, учитывать эти особенности при использовании баз данных поисковых систем как материала для лингвистических исследований.

Важно, какую информацию и в каком виде можно извлечь из выходных интерфейсов информационной поисковой системы. Из всех реквизитов на странице с результатами поиска наибольший интерес для задач лингвистического исследования представляют частотные характеристики и выдача контекста. Следует различать два типа частот,ываемых и выдаваемых системами, — пословную и подокументную. 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 [Захаров, 2003; 2005].

Однако статистические результаты работы поисковых систем всегда приблизительны, по ним можно составить лишь общее впечатление о том, как обстоят дела в языке с той или иной лексической единицей. В целом использование поисковых систем как корпусных инструментов связано со многими проблемами [Беликов, Селегей, Шаров, 2012, с. 42–44].

Существовали также проекты специальных поисковых систем-посредников, имеющих корпусный интерфейс, но пользующихся базами данных поисковых систем. С 1998 г. разрабатывается инструмент WebCorp, позволяющий лингвистам получать лингвистические результаты из Интернета (<http://wse1.webcorp.org.uk>) [Renouf, Kehoe, Banerjee, 2006]. Однако в целом этот путь оказался малопродуктивным и периферийным.

## 6.2. Веб как корпус

Затем возникла идея создания полноценных корпусов, функционирующих под управлением корпусных менеджеров, на основе текстов, взятых из Интернета. Особенно активно эта проблема стала обсуждаться после доклада Адама Килгарриффа в 2001 г. [Kilgarriff, 2001]. Затем она была развита в других работах [Kilgarriff, Grefenstette, 2003; Baroni, Bernardini, Ferraresi et al., 2009].

В начале 2000-х гг. было организовано сообщество лингвистов и специалистов по информационным технологиям под названием WaCky (The Web-As-Corpus Kool Yinitiative, <https://wacky.sslmit.unibo.it/doku.php>). Его участники стали разрабатывать набор инструментов (и интерфейсов для существующих инструментов), позволяющих лингвистам искать в Интернете тексты (веб-страницы) для наполнения корпусов, обрабатывать их, индексировать и в конечном итоге создавать из них корпусы, которые могут насчитывать миллиарды токенов. Эта технология и эти корпусы получили название WaC (bootstrap specialized corpora and terms from the web — BootCaT) (см. материалы семинаров WaC по адресу <https://sigwac.org.uk/>). И в течение 2006–2009 гг. были созданы и предоставлены в открытый доступ корпусы английского, немецкого, французского и итальянского языков (ukWaC, deWaC, frWaC, itWaC) объемом 1–2 млрд токенов [Baroni, Bernardini, Ferraresi et al., 2009]. В 2011 г. в Свободном университете в Берлине (Freie Universität) стартовал проект COW (COrpora from the Web). В его рамках были созда-

ны английский, немецкий, французский, голландский, испанский и шведский корпусы [Schäfer, Bildhauer, 2012]. К 2014 г. размер некоторых из них достиг 10 млрд токенов. Немецкий корпус насчитывал 20 млрд токенов [Schäfer 2015]. Эти корпусы доступны через веб-портал проекта (<https://webcorpora.org/>). На сайте также представлены частотные словари английского, немецкого, испанского и шведского языков, сформированные на базе веб-корпусов. Большое количество веб-корпусов было создано в рамках проекта CLARIN в Словении в Институте Йожефа Стефана (Jožef Stefan Institute). Помимо южнославянских корпусов (baWaC, hrWaC, slWaC, srWaC) [Ljubešić, Erjavec, 2011; Ljubešić, Klubička, 2014], там были созданы корпусы и для многих других языков, включая японский. Их размер варьирует от 400 млн до 2 млрд токенов. Большинство корпусов доступны через интерфейс системы NoSketch Engine без каких-либо ограничений.

Наибольшее количество веб-корпусов создано компанией Lexical Computing Ltd. (Брайтон, Великобритания; Брно, Чешская Республика), которые доступны в среде Sketch Engine [Jakubíček, Kolgar, Kovář et al., 2013]. На сайте компании насчитывается несколько сотен таких корпусов почти для 100 языков мира, и их размеры варьируются от нескольких сотен тыс. токенов (игбо, татарский, тайский, мальдивский языки), нескольких млн токенов (идиш,yoruba, самоа) до 16 млрд токенов (английский, немецкий, русский). Размер крупнейшего корпуса русского языка составляет 14,5 млрд токенов.

Очевидно, что ни один корпус не может сравниться по репрезентативности языкового материала с вебом. Однако исследования, проведенные на основе WaC-корпусов, выявили проблемы, которые можно разделить на три группы: проблемы лингвистической разметки, проблемы метаразметки и технические проблемы, связанные с удалением дублей, элементов гипертекстовых языков разметки веб-документов и т. п.

Проблемы лингвистической разметки заключаются в необходимости размечать тексты, более сложные по языку с точки зрения лексического состава и синтаксической структуры, чем тексты традиционных корпусов, и гораздо более низкого качества. В текстах веб-документов встречается много орфографических ошибок, гораздо больше вариативность различного типа, многие слова оказываются разорванными из-за знаков переноса и т. п.

Проблемы, которые мы условно назвали проблемами метаразметки, в отличие от традиционных корпусов, выглядят перевернутыми с ног на голову. Работая по традиционной технологии, мы имеем дело с документами, которые надо разметить. В соответствии с новой технологией перед нами стоит задача отобрать документы, соответствующие заданным элементам метаразметки. При этом нужно учесть, что полноценная традиционная метаразметка (например, в терминах TEI) по отношению к веб-документам неприменима по причине ее отсутствия в веб-документах. В настоящее время на практике мы получаем корпусы с минимальной метаразметкой в терминах веба (домен или доменное имя, дата помещения на сайт или дата формирования корпуса, длина документа и др.), и, следовательно, ничего нельзя сказать об их сбалансированности в терминологии традиционных корпусов. То есть мы получаем корпусы большого объема, но при этом возникает вопрос их качества.

Тем не менее за прошедшее десятилетие технология WaC достигла заметных успехов и нашла широкое практическое применение. Создан и продолжает совершенствоваться соответствующий набор прикладных программ, которые поддерживают эффективную реализацию этой технологии, включая инструменты для сканирования веб-страниц, распознавания языка, очистки данных, дедупликации, причем многие из них предоставляются бесплатно или с открытым кодом.

В 2018 г. был запущен в свободную эксплуатацию сервис Boot-Cat <https://bootcat.dipintra.it/> [Baroni, Bernardini, 2004], позволяющий любому пользователю создавать свои корпусы на основе текстов из веба. На первом этапе пользователи задают список одно- или многословных терминов, на их основе формируются поисковые запросы различной длины, которые отправляются в поисковую систему, возвращающую результаты поиска в вебе в виде списка потенциально релевантных URL-адресов. На этом этапе пользователь имеет возможность проверять URL-адреса и удалять нерелевантные. Затем веб-страницы с этих адресов (или с оставшихся после удаления нерелевантных) загружаются, преобразуются в обычный текст, из них удаляется нерелевантная (нетекстовая) информация, и они сохраняются как файлы в формате «txt». Далее эти тексты могут быть загружены в тот или другой конкордансер или корпусный менеджер. Используя BootCat в таком режиме, можно быстро создать относительно большой корпус (обычно около 80 текстов с параметрами по

умолчанию и без ручного контроля качества, поэтому с определенным «шумом») менее чем за полчаса. В целом программно-технологическая цепочка поддерживает различные режимы и может быть использована для создания и более крупных и качественных корпусов.

Новые подходы к созданию корпусов на базе веба были заявлены и реализуются и в России в проекте под названием «Генеральный интернет-корпус русского языка» [Беликов, Копылов, Пиперски, и др., 2013; Пиперски, 2013; Шаров, Беликов, Копылов и др., 2015]. С одной стороны, это веб-корпус, с другой стороны, он создается путем целевого отбора определенных «зон» в интернет-пространстве, а не более или менее случайного отбора веб-документов.

Интересный проект большого веб-корпуса (более 5 млрд токенов), предназначенного для задач машинного обучения, представляет собой морфологически и синтаксически размеченный корпус Taiga, доступный для выгрузки под лицензией CC BY-SA 3.0 ([https://tatianashavrina.github.io/taiga\\_site/](https://tatianashavrina.github.io/taiga_site/)) [Shavrina, Shapovalova, 2017].

### 6.3. Технология WaC

Чтобы создать веб-корпус, обычно необходимо последовательно выполнить следующие операции:

- загрузка больших объемов данных из Интернета;
- извлечение из них текстовой информации;
- нормализация кодирования, если требуется;
- идентификация языка загруженных текстов, удаление «неправильных» документов;
- сегментирование текста на абзацы и предложения;
- удаление дублей на уровне документов;
- удаление дублей на уровне сегментов;
- токенизация;
- лингвистическая (морфологическая и, возможно, синтаксическая) разметка;
- загрузка полученного корпуса в корпусный менеджер.

За исключением первых двух, все остальные операции уже присутствовали (в определенной степени) в процессе создания традиционных корпусов. Поэтому в случае необходимости можно вос-

пользоваться существующими инструментами и методологией, в первую очередь если это касается лингвистической разметки.

Загрузка данных из Интернета обычно выполняется с помощью одной из двух стандартных методологий, которые по-разному извлекают URL-адреса загружаемых веб-страниц.

В рамках метода, описанного С.А. Шаровым [Sharoff, 2006], список среднечастотных слов используется для генерации случайных кортежей длины  $n$ , которые передаются в поисковой системе. Верхняя часть списка найденных URL-адресов (предположительно самых релевантных) используется для выгрузки данных для корпуса. Процесс может быть частично автоматизирован программой BootCaT [Baroni, Bernardini, 2004].

Второй метод основан на сканировании веб-пространства с помощью специальной программы, которая использует исходный список веб-адресов, предоставленных пользователем, и итеративно ищет новые URL-адреса, анализируя гиперссылки в уже загруженных веб-страницах. Программа обычно работает автономно и способна выполнять идентификацию языка и/или дедупликацию «на лету», что делает весь процесс более эффективным и позволяет относительно быстро (в течение нескольких часов или дней) загружать текстовые данные, содержащие несколько сотен миллионов токенов. Две самые популярные программы, используемые для обхода веб-пространства, — это универсальный робот-поисковик Hiritrix (<https://webarchive.jira.com/wiki/display/Heritrix>) и специализированный лингвистический инструмент SpiderLing [Suchomel, Pomikálek, 2012].

Каждый из упомянутых выше методов имеет свои плюсы и минусы, причем первый из них более подходит для создания небольших корпусов (особенно если корпус ориентирован на конкретную предметную область), в то время как последний обычно используется для создания очень больших корпусов объемом в несколько миллиардов токенов.

Описан опыт создания корпусов семейства Aranea по технологии WaC [Benko, 2014; Benko, Zakharov, 2016]. В последней работе представлен опыт создания сверхбольшого корпуса серии Maximum для русского языка объемом 13,7 млрд токенов (со времени написания статьи корпус «подрос» и сейчас насчитывает 19,8 млрд токенов). Суть этой технологии заключается в объединении двух вышеописанных стандартных методологий извлечения данных для корпуса из веба. Вначале с помощью BootCaT путем нескольких итераций

были собраны URL-адреса (более 200 тысяч), в совокупности отражающие весь лексический спектр русского языка. Затем эти адреса итерационно подавались на вход программы SpiderLing. Этот процесс требует мощной вычислительной техники и занимает несколько месяцев непрерывной работы вычислительного комплекса. Приводятся объемные и временные характеристики процесса создания корпуса Araneum Russicum Maximum [Benko, Zakharov, 2016, 90–91].

## Глава 7. Обзор существующих корпусов различных типов

В настоящее время существуют национальные общеязыковые корпусы для большинства основных языков мира. Языки, не имеющие национального корпуса в своей стране, тем не менее часто тоже представлены на том или ином сайте. Число разных корпусов измеряется тысячами и постоянно растет. М. В. Копотев пишет, что два крупнейших специализированных каталога CLARIN ([www.clarin.eu/](http://www.clarin.eu/)) и ELRA (<http://www.elra.info/>) содержат информацию о более чем 3000 корпусах [Копотев, 2014]. Список лингвистических ресурсов, включающий и корпусы, можно найти на сайте Ассоциации компьютерной лингвистики по адресу: [### 7.1. Зарубежные корпусы](https://aclweb.org/aclwiki>List_of_resources_by_language</a>.</p></div><div data-bbox=)

Приведем некоторые национальные и другие корпусы с указанием объемных характеристик (табл. 7.1).

Таблица 7.1. Список корпусов с указанием объемных характеристик

Корпус	Название и адрес в сети	Объем (с/у)
Корпус американского варианта английского языка	Corpus of Contemporary American English (COCA) <a href="https://corpus.byu.edu/coca/">https://corpus.byu.edu/coca/</a>	560 млн
Корпус английского языка	British National Corpus (BNC) <a href="http://corpus.byu.edu/">http://corpus.byu.edu/</a> или <a href="http://sara.natcorp.ox.ac.uk/">http://sara.natcorp.ox.ac.uk/</a>	100 млн

Продолжение табл. 7.1

Корпус	Название и адрес в сети	Объем (с/у)
Корпус арабского языка	arabiCorpus <a href="http://arabicorpus.byu.edu/">http://arabicorpus.byu.edu/</a>	174 млн
Корпус арабского языка (язык Корана)	The Quranic Arabic Corpus (Коран) <a href="http://corpus.quran.com/">http://corpus.quran.com/</a> 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 <a href="http://dcl.bas.bg/bulnc/">http://dcl.bas.bg/bulnc/</a>	1,2 млрд
Корпус венгерского языка	Венгерский национальный корпус <a href="http://mnsz.nytud.hu/index_hun.html">http://mnsz.nytud.hu/index_hun.html</a>	188 млн
Корпус датского языка	Корпус датского языка KorpusDK <a href="http://ordnet.dk/korpusdk">http://ordnet.dk/korpusdk</a>	56 млн
Корпус испанского языка	Корпус испанского языка (проект М. Дэвиса) <a href="http://www.corpusdelespanol.org/">http://www.corpusdelespanol.org/</a>	100 млн
Корпус испанского языка	Corpus de Referencia del Español Actual (CREA) <a href="http://corpus.rae.es/creanet.html">http://corpus.rae.es/creanet.html</a>	150 млн
Корпус итальянского языка	Корпус итальянских текстов Болонского университета CORIS <a href="http://corpora.dslo.unibo.it/">http://corpora.dslo.unibo.it/</a>	130 млн
Корпус китайского языка	The LIVAC Synchronous Corpus (Linguistic Variations in Chinese Speech Communities) <a href="http://www.livac.org/">http://www.livac.org/</a>	600 млн
Корпус китайского языка	Scripta Sinica database (база данных текстов) <a href="http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm">http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm</a>	683 млн
Корпус немецкого языка	Немецкий справочный корпус Das Deutsche Referenzkorpus (DeReKo) <a href="http://www.ids-mannheim.de/kl/projekte/korpora/">http://www.ids-mannheim.de/kl/projekte/korpora/</a>	5,4 млрд

Окончание табл. 7.1

Корпус	Название и адрес в сети	Объем (с/у)
Корпус немецкого языка	Синтаксически аннотированный корпус немецкого языка NEGRA <a href="http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus">http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus</a>	355 тыс., (20600 предл.)
Корпус польского языка	PELCRA Reference Corpus of Polish <a href="http://pelcra.pl/3-2?lang=pl">http://pelcra.pl/3-2?lang=pl</a>	100 млн
Корпус польского языка	Корпус польского языка IPI PAN <a href="http://korpus.pl/index.php?lang=pl&amp;page&gt;Welcome">http://korpus.pl/index.php?lang=pl&amp;page&gt;Welcome</a>	250 млн
Корпус словацкого языка	Словацкий национальный корпус <a href="http://korpus.juls.savba.sk/">http://korpus.juls.savba.sk/</a>	1,2 млрд
Корпус словенского языка	Корпус словенского языка FidaPLUS <a href="https://old.sketchengine.co.uk/corpus/first_form?corpname=preloaded/fidaplus2">https://old.sketchengine.co.uk/corpus/first_form?corpname=preloaded/fidaplus2</a>	600 млн
Корпус словенского языка	Nova beseda <a href="http://bos.zrc-sazu.si/a_beseda.html">http://bos.zrc-sazu.si/a_beseda.html</a>	318 млн
Корпус французского языка	American and French Research on the Treasury of the French Language (ARTFL-FRANTEXT) <a href="http://artfl-project.uchicago.edu/content/artfl-frantext">http://artfl-project.uchicago.edu/content/artfl-frantext</a>	215 млн
Корпус французского языка	Lexiquum <a href="http://retour.iro.umontreal.ca/cgi-bin/lexiquum">http://retour.iro.umontreal.ca/cgi-bin/lexiquum</a>	229 млн
Корпус чешского языка	Чешский национальный корпус <a href="http://korpus.cz/">http://korpus.cz/</a>	1,3 млн
Корпус шведского языка	Банк шведского языка (разные корпусы и словари) <a href="http://spraakbanken.gu.se/">http://spraakbanken.gu.se/</a>	1,3 млрд
Корпус японского языка	The Balanced Corpus of Contemporary Written Japanese (BCCWJ) <a href="https://pj.ninjal.ac.jp/corpus_center/bccwj/en/">https://pj.ninjal.ac.jp/corpus_center/bccwj/en/</a>	100 млн

Остановимся подробнее на некоторых корпусах.

**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 содержит 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 содержит тексты разных стилей и не ограничен по тематике. Подкорпус устной речи представляет речь добровольно вызвавшихся участвовать в проекте людей различных возрастов, проживающих в разных частях Великобритании и принадлежащих к различным социальным классам. Разговорная речь присутствует в самых разных контекстах — от разговоров на формальных деловых или правительственные встречах до радиошоу и телефонных разговоров.

Все тексты BNC сегментированы по предложениям. Словам внутри предложения присвоены соответствующие маркеры, обозначающие грамматический класс слова или его часть речи. Знакам препинания тоже присвоены соответствующие маркеры. Сегментацию и автоматическое присвоение словам тегов выполнила программа CLAWS, разработанная в Ланкастерском университете. Ошибочная разметка не превышает 1,7 %. Кроме того, если программа автоматической разметки не могла однозначно присвоить слову какой-то маркер, присваивал ему сразу два маркера (например, VVD и VVN — первый обозначает глагол прошедшего времени, а второй — причастие прошедшего времени). Такие «синонимичные» маркеры имеют примерно 4,7 % слов 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 BNC, отбирались и балансировались по трем основным критериям: время, область, которую данный текст описывает, и тип издания. По времени все тексты принадлежат примерно к одному периоду начиная с 1975 г., исключения дела-

лись только для художественной литературы, поскольку некоторые произведения более раннего периода популярны и по сей день. К области художественной литературы относится 25 % текстов. В BNC представлены литературные произведения, написанные не ранее 1964 г. 75 % письменных текстов было взято 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) *контекстно-управляемую* часть, содержащую записи, сделанные на каких-либо публичных мероприятиях, где важную роль играл контекст.

1) *Демографическая* часть. В записи диалогов участвовали 124 добровольца, живущих на всей территории Великобритании. Они должны были носить с собой магнитофоны в течение нескольких дней, с тем чтобы при выполнении различных действий фиксировать в записных книжках, в каких условиях состоялись разговоры, кто являлся собеседниками, каковы были их взаимоотношения, физическое окружение в момент записи речи и т. д. Добровольцы отбирались так, чтобы было примерно равное количество мужчин и женщин из каждой 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, а тексты записаны обычной английской орфографией. Эти разговоры сейчас используются как основа изучения характера устной речи, и результаты оказываются полезными и интересными [Finegan, 2004].

2) *Контекстно-управляемая* часть. Создатели преследовали цель собрать равное количество записей, относящихся к следующим четырем довольно широким категориям социального контекста:

- образовательные и информативные собрания: лекции, программы новостей, обсуждение чего-либо в классе, семинары;
- деловые события: выставки, консультации, интервью, собрания торговых организаций;

- публичные события: проповедь, политические речи, заседания парламента;
- темы, касающиеся досуга: спортивные комментарии, клубные встречи.

На основе разметки SGML разработчики создали собственный корпусный менеджер, который назвали SARA (SGML Aware Retrieval Application). SARA был изначально разработан как программа «клиент/сервер», то есть система, где один компьютер или более имеет сетевой 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* — ЧНК). Это синхронный морфологически размеченный корпус, представляющий современный чешский язык. Созданием корпуса занимается Институт ЧНК Карлова университета в Праге. Институт был создан в 1994 г. и функционирует на средства грантов, спонсоров и при поддержке Министерства образования Чешской Республики. С информацией об институте можно ознакомиться на сайте <https://ucnklff.cuni.cz/>, а сам корпус доступен по адресу: <http://korpus.cz>.

Первый корпус чешского языка был создан в 1999 г. 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: художественные тексты составляют 15 % ЧНК, в то время как остальные 85 % представлены информативными текстами (публицистическими (60 %) и научными (25 %), относящимися к социальным, естественным, техническим наукам, искусствоведению и т. п.).

Первоначальный корпус, насчитывавший 100 млн словоупотреблений письменных текстов, содержал также небольшие коллекции разговорной 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, SYN2010 и SYN2015, а также большое число кор-

пусов различных типов. Все синхронные корпусы объединены в общий пул объемом 4,5 млрд словоупотреблений. Кроме того, в составе ЧНК создано несколько корпусов устной (разговорной) речи общим объемом 6 млн словоупотреблений, диахронический корпус (4 млн словоупотреблений), параллельный корпус InterCorp (более чем 30 языков, 1,7 млрд словоупотреблений), публицистический корпус (1 млрд слов) и др. (см.: <https://wiki.korpus.cz/doku.php/cnk:uvod>).

Работа с корпусом осуществляется через интерфейс KONTEXT, построенный над корпусным менеджером NoSketch Engine. Большой интерес представляют дополнительные инструменты, работающие на базе корпусов: SyD (сравнение различных статистических характеристик для пары заданных слов в синхронном и диахроническом корпусах или в корпусах письменной и устной речи), Morfio (исследование продуктивности словообразовательных моделей в чешском языке на базе корпусных данных), Kworts (выделение ключевых слов из заданного текста), Treq (формирование словарей переводных эквивалентов с конкордансом и частотными характеристиками на базе параллельных корпусов).

**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) в университете Бригама Янга (Brigham Young University) в 2008 г. В феврале 2019 г. объем корпуса COCA, включающего тексты с 1990 по 2017 г., равномерно представляющие устную речь, художественную прозу, популярные журналы, газеты и научную литературу, составил 560 млн слов. Он обновляется два раза в год и удобен для наблюдения за текущими изменениями, происходящими в языке.

Кроме того, по указанному адресу <http://corpus.byu.edu/> находится много других корпусов. Среди них стоит отметить корпусы Google Books, Global Web-Based English (веб-корпус английского языка из 20 англоговорящих регионов) и новый The Intelligent Web-based Corpus (US/CA/UK/IE/AU/NZ) объемом 14 млрд слов. Особенности корпуса описаны в файле по адресу [https://corpus.byu.edu/iweb/help/iweb\\_overview.pdf](https://corpus.byu.edu/iweb/help/iweb_overview.pdf). Еще одна важная особенность указанной службы в том, что это один из самых функционально мощных корпусных менеджеров.

Следует упомянуть крупный и оригинальный корпусный проект — **диахронический корпус Ngram Viewer**, созданный на основе библиотеки Google Books (<https://books.google.com/ngrams>) [Michel, 2011]. Сейчас это наиболее мощный инструмент для диахронических исследований, который содержит огромные корпусы размеченных текстов книг на 9 языках. Например, корпус книг на русском языке содержит 591 310 текстов общим объемом более 67 млрд словоупотреблений [Захаров, Масевич, 2014]. Самые поздние публикации, включенные в систему, относятся к 2008 г.

В 2011 г. часть корпуса, а именно Google Books (American English) Corpus, объемом 155 млрд слов, основанная на данных Google Books и включающая тексты книг на американском варианте английского языка с 1810 по 2009 гг., была размещена на корпусном сайте университета Бригама Янга (<http://corpus.byu.edu>). Затем к нему прибавилось подмножество английского и испанский корпус из Google Books Ngram Viewer. Все они доступны под интерфейсом Corpus/byu.edu.

Из немецких корпусов необходимо упомянуть о **Корпусе немецкого языка DeReKo** (das Deutsche Referenz Korpus), доступном по адресу <http://www.ids-mannheim.de/kl/projekte/korpora/>. Электронное собрание текстов, созданное в рамках корпусного проекта Института немецкого языка в Мангейме (Германия), состоит из беллетристики, научных и публицистических текстов и содержит более 42 млрд словоупотреблений (по данным на ноябрь 2018 г.). Этот корпус оформлен как собрание отдельных немецкоязычных подкорпусов. Корпус содержит морфосинтаксическую разметку, разработанную в соответствии с рекомендациями TEI. Однако размечены далеко не все подкорпусы. Корпусный менеджер COSMAS II, обеспечивающий работу с корпусом, позволяет осуществлять поиск по лексическим единицам и по морфологическим признакам словоформ. Также используются логические операторы, позиционные операторы, операторы расстояния и др.

Заслуживает упоминания **Корпус немецкого языка**, являющийся частью лексической системы DWDS (Digitales Wörterbuch der deutschen Sprache), объединяющей различные корпусы, текстовую базу данных (13 млрд единиц) и словарь немецкого языка (465 тыс. слов). Объем корпуса, доступного всем пользователям, составляет 5,5 млрд токенов. Корпус был создан Берлин-Бранденбургской академией наук и работает под управлением корпусного менеджера DDC. Он включает в себя несколько подкорпусов:

- художественная литература (26 %);
- газеты (27 %);
- научная литература (22 %);
- нехудожественные тексты (20 %);
- устные тексты (5 %).

Корпус размечен морфологически и семантически (на уровне слов), также размечены предложные конструкции и именные группы. Корпус соединен со словарем немецкого языка. В корпусе можно выбирать временные периоды и типы текстов. Являющийся его подкорпусом **Немецкий национальный корпус** представляет собой 100-миллионный аналог Британского национального корпуса, покрывающий весь ХХ в. (1900–2000 гг.). Для работы с основным корпусом требуется регистрация.

Наибольшее количество корпусов собрано на сайте компании Lexical Computing Ltd. (Брайтон, Великобритания; Брно, Чешская Республика), которые загружены и поддерживаются системой **Sketch Engine** (<http://sketchengine.co.uk/>). Нужно сказать, что это мощная корпусная служба, которая предоставляет зарегистрированным пользователям как услуги по созданию корпусов и различные сервисные функции, так и возможность пользоваться всей совокупностью корпусов, созданных в рамках службы и представленных на сайте системы. По состоянию на август 2019 г. число таких корпусов составляет более 500 и охватывает 92 языка. Среди них имеется ряд корпусов русского языка, прежде всего корпус объемом 14,5 млрд словоупотреблений, созданный из текстов Интернета по технологии WaC. В 2018 г. появился новый интерфейс для работы с системой, ориентированный на мобильные устройства. По количеству и объему корпусов и по своим функциональным возможностям система Sketch Engine является самой мощной системой в мире корпусной лингвистики [Kilgarriff Baisa, Bušta et al., 2014].

Менеджер системы 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 (подробнее см. л. 8.4, 8.6).

**Корпусы семейства Aranea** (лат. паутина) созданы в рамках совместного проекта Братиславского университета Коменского и Сло-

вацкой академии наук. В настоящее время это семейство состоит из сопоставимых веб-корпусов, созданных по технологии WaC для 22 языков (см. <http://unesco.uniba.sk/guest/index.html>). Корпусы для некоторых других языков находятся в процессе подготовки. Важно подчеркнуть единство технологических решений при создании корпусов, приблизительную одновременность их создания, унификацию результатов разметки для разных языков. Корпусы, как правило, представлены двумя типоразмерами: серия Maius («большие») объемом 1,2 млрд токенов, или около 1 млрд слов, — токены, начинающиеся с буквенных символов, и серия Minus («малые») объемом 120 млн токенов, или около 100 млн слов. Каждый корпус этой серии представляет собой 10 % случайную выборку соответствующего корпуса Maius.

Для некоторых языков существуют также региональные варианты. Так, Araneum Russicum Maius & Minus содержат русские тексты, загруженные с любых интернет-доменов, Araneum Russicum Russicum Maius & Minus содержат тексты, загруженные с доменов «.ru», «.su» и «.рф», а Araneum Russicum Externum Maius & Minus основаны на текстах «нерусских» доменов, таких как «.ua», «.by», «.kz» и др. Точно так же имеется шесть корпусов для английского языка — Araneum Anglicum, Araneum Anglicum Africanum, Araneum Anglicum Asiaticum, каждый в двух типоразмерах, и шесть для французского — Araneum Francogallicum, Araneum Francogallicum Africanum, Araneum Francogallicum Canadiense, каждый в двух типоразмерах. Кроме того, имеются сверхбольшие корпусы серии Maximum для следующих языков: русский (19,8 млрд токенов), английский (11,4 млрд токенов), французский (8,7 млрд токенов), чешский (5,17 млрд токенов), два корпуса словацкого языка (4,34 млрд и 2,96 млрд токенов), немецкий (9,09 млрд).

Корпусы загружены в корпусный менеджер NoSketch Engine и доступны для использования по адресу <http://unesco.uniba.sk>. Более подробную информацию о проекте Aranea см. в работах В. Бенко, В. П. Захарова [Benko, 2014; Benko, Zakharov, 2016].

## 7.2. Корпусы русского языка

### 7.2.1. Первые корпусы русского языка

Первый русскоязычный корпус был создан в 1980-е годы в Университете Уппсалы (Швеция), послужил основой для частотного словаря русского языка [Lönngrén, 1993]. Однако еще до первых русскоязычных корпусов в 1960–1970-е годы был создан **Частотный словарь русского языка** [Засорина, 1977], который де-факто был сформирован на базе корпуса объемом в 1 млн словоупотреблений. Корпус включал общественно-политические тексты, художественную литературу, научные и научно-популярные тексты из разных областей и драматургию (как своего рода аналог устной речи) примерно в равной пропорции.

Инициатором и руководителем проекта была Л. Н. Засорина, ею же были разработаны теоретические основы и практическая инструкция обработки лексического материала. В результате была создана аналитическая модель переработки сегментов текста в элементы словаря, которую можно назвать аналитической грамматикой русского языка (отдельно для каждой части речи) и которая применялась при предмашинной обработке текста. Знакомство с методами создания словаря показывает, что уже тогда и там обсуждались все вопросы, которые сейчас обсуждаются в корпусной лингвистике. В частности, это проблема репрезентативности и сбалансированности корпуса, или жанровой дифференциации лингвистического ресурса.

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

В 1985 г. в СССР по инициативе академика А. П. Ершова были начаты работы по созданию **Машинного фонда русского языка** [Машинный фонд 1989]. В создании фонда принимали участие более 40 организаций-соисполнителей, среди них — Институт русского языка, Московский, Ленинградский, Харьковский, Гродненский, Сыктывкарский и Саратовский университеты и др. В задачи фонда

входило накопление на машинных носителях текстовых, лексикографических и грамматических источников, необходимых для научного изучения русского языка и для осуществления прикладных разработок в виде корпусов и баз данных. Одновременно создавались программные средства для проведения лингвистических исследований [Национальный корпус русского языка]. В 1985–1992 гг. разработаны концепция и архитектура Машинного фонда русского языка, концепция терминологического банка данных. Тогда же стали вводить в компьютер словари и академическую грамматику русского языка, сформировали машиночитаемые тексты по русской литературе XIX–XX вв., а именно корпусы поэзии и художественной прозы, а также корпусы 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** 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 — представить в первую очередь литературный язык, поэтому в массиве нет образцов разговорной речи.

В корпус отбирались специальные тексты (включались не фрагменты, 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 удалено более важным, с точки зрения создателей корпуса, темам, а среди художественных текстов предпочтение отдано более известным авторам. Тексты в корпусе записывались латиницей и специальными знаками. 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 ot-novenie k nim*

Уппсальский корпус вошел в так называемые **Тюбингенские корпусы русских текстов**, созданные в 1990–2000-е годы силами специального научно-исследовательского сектора SFB 441 Тюбингенского университета с возможностью поиска онлайн (<http://www.lingexp.uni-tuebingen.de/sfb441/b1/rus/korpora.html>).

Корпусы размечены тегами морфологической аннотации. Разметка была осуществлена при помощи статистического теггера (TnT). Поиск может производиться как по словоформам, так и по морфологическим тегам. Возможен вывод текста вместе с разметкой. Для ввода поискового выражения и вывода найденного текста можно выбрать одну из кодировок: KOI8 или кириллицу Windows-1251 — либо 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 под руководством А. А. Поликарпова ([http://www.philol.msu.ru/~lex/corpus/corp\\_descr.html](http://www.philol.msu.ru/~lex/corpus/corp_descr.html)). 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 в период 1994–1997 гг. (23 110 текстов), представленности в нем ежедневных и неежедневных газет, левых и правых, центральных и местных, общих и профессионально ориентированных газет. 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. п. 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: развлекательные жанры (игра, кроссворд, гороскоп и т. д.), жанр религиозной проповеди, жанры, которые пока трудно отнести к определенному типу.

### ***7.2.2. Современные корпусы русского языка***

#### ***7.2.2.1. Национальный корпус русского языка***

Достаточно 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 уже в 1980-х годах [Сичинава, 2005].

**Национальный корпус русского языка** (НКРЯ) был создан в начале XXI в. и впервые размещен в Сети 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 критериюreprезентативности и другим требованиям, предъявляемым к современным корпусам, о чём свидетельствуют его характеристики (см. статистику на сайте <http://ruscorpora.ru/corpora-stat.html>):

- объем, который в сумме составляет около 676 млн словоупотреблений (по данным сайта на ноябрь 2018 г.);
- жанровое разнообразие текстов, которые относятся ко всем основным сферам использования русского языка (научной, официально-деловой, публицистической, церковно-богословской, художественной, разговорно-бытовой, включая устную и электронную коммуникацию);
- 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.);
- наличие текстов, относящихся к разным периодам создания, что позволяет проследить изменения в употреблении языковых явлений и, возможно, установить динамику этих изменений [Гришина, Савчук, 2008].

Объем основного корпуса составляет 288,7 млн словоупотреблений. 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. В корпусе можно условно выделить две части — современную и диахроническую. Корпус современных текстов составляют тексты, период создания которых укладывается в рамки 1951–2010 гг. Диахроническая часть объединяет тексты XVIII, XIX и первой половины XX в.

Объем корпуса позволяет изучать вариативность и изменчивость достаточно частотных языковых явлений, а также получать надежные результаты по следующим направлениям:

- изучение морфологических вариантов имен, глаголов и других частей речи и их эволюции;
- исследование словообразовательных вариантов и связанный с ними проблемы паронимов, продуктивности словообразовательных моделей и словообразовательных средств;
- исследование изменения вариантов управления, согласования и примыкания;
- исследование акцентологических вариантов и изменений в акцентной системе русского языка;
- исследование лексической вариативности, в частности изменения состава синонимических рядов и тематических групп, а также семантических соотношений в них [Гришина, Савчук, 2008].

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

- газетный корпус, охватывающий статьи из средств массовой информации 2000-х годов;
- глубоко аннотированный (синтаксический) корпус, содержащий тексты, снабженные морфосинтаксической разметкой, где помимо морфологической информации, приписанной каждому слову текста, для каждого предложения задана его синтаксическая структура (дерево зависимостей);
- совокупность параллельных двуязычных корпусов текстов на разных языках (английском, французском, немецком, испанском, итальянском, польском, украинском, белорусском и др.);
- 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, например поиск 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 о русском

ударении; 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, который включает расшифровки магнитофонных записей публичной и частной устной речи, а также транскрипты кинофильмов 1930–2000-х годов (подробнее см. в п. 7.2.2.5).
- мультимедийный русский корпус (МУРКО), образованный фрагментами кинофильмов 1930–2000-х годов, представленными 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 типу речевого действия (согласие, ирония). В поисковой выдаче видеофрагменты доступны для просмотра и прослушивания;
- исторический корпус.

#### 7.2.2.2. Хельсинкский аннотированный корпус (ХАНКО)

**Корпус ХАНКО** (<http://www.ling.helsinki.fi/projects/hanco/>) создан в Хельсинкском университете в начале 2000-х годов как часть проекта «Функциональный синтаксис русского языка» (рук. — проф. А. Мустайоки) и постоянно развивается. 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. Особенности корпуса — тщательно проработанный формат лингвистического описания данных и полная визуальная проверка результатов автоматической разметки, обеспечившая полное снятие грамматической омонимии. Кроме того, размечены многословные устойчивые обороты (примерно 2000 единиц). Синтаксическая разметка корпуса представляет собой разметку в терминах членов предложения, планируется разметка в терминах деревьев зависимостей, семантическая разметка в терминах семантических категорий.

### *7.2.2.3. Корпусы университета г. Лидс*

В 2000-е годы в университете г. Лидс, в Центре переводческих исследований, С. А. Шаровым создано большое количество корпусов для разных языков (английского, арабского, китайского, французского, немецкого, итальянского, японского, испанского, польского и др.) (<http://corpus.leeds.ac.uk/>). Среди них имеются корпусы и русского языка (<http://corpus.leeds.ac.uk/ruscorpura.html>). Это версия Национального корпуса русского языка объемом в 116 млн словоупотреблений (на ее основе 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 вышенназванных.

Поисковый интерфейс Leeds CQP базируется на корпусном менеджере IMS Corpus Workbench и предоставляет интересные возможности. Он позволяет вести очень точный лексико-грамматический поиск: можно использовать специальный язык запросов, в том числе с применением языка регулярных выражений. Предусмотрены 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/>). Одновременно эти корпусы доступны через новый, более развитый корпусный менеджер IntelliText (<http://corpus.leeds.ac.uk/it/>).

### *7.2.2.4. 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. В нем можно осуществлять поиск 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.

#### *7.2.2.5. Устные корпусы русского языка*

Для целенаправленного изучения особенностей устной речи можно воспользоваться специально созданными корпусами устной русской речи.

По состоянию на июль 2019 г. устный подкорпус Национального корпуса русского языка насчитывает 12,1 млн словоупотреблений и включает в себя расшифровки магнитофонных записей публичной и частной устной речи, а также транскрипты кинофильмов. Использована русская стандартная орфография (при этом приводятся наиболее частотные и общепринятые стяженные формы). В корпусе возможен лексический, морфологический и семантический поиск, формирование пользовательских подкорпусов, в том числе и по социологическим параметрам. Включены тексты самых разных жанров и типов, разного происхождения с точки зрения географии (Москва, Санкт-Петербург, Саратов, Ульяновск, Таганрог, Екатеринбург, Норильск, Воронеж, Новосибирск и мн. др.) [Национальный корпус русского языка]. 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 (0,23 %), художественное чтение — 2,5 тыс. (0,024 %), театральная речь — 4,3 тыс. (0,041 %). Мультимедийный корпус включает прежде всего речь кино — 3,4 млн (32,8 %), публичную речь — 31,5 тыс. (0,3 %), непубличную речь — 12,6 тыс. (0,12 %) и др.

Нормативную русскую речь представляет Корпус транскрибированных русских устных текстов (Корпус русских спонтанных текстов), который до сих пор остается единственным общедоступным корпусом русской речи, снабженным полной фонетической транскрипцией (<http://narusco.ru/search/trn-search.php>). Основная цель создания корпуса — его последующее применение для моделирования восприятия естественной звучащей речи [Риехакайнен, 2917].

В качестве примера корпусов, наполненных устными текстами, особенно относящимися к непубличной речи — телефонным разговорам, неформальным беседам и т. д., рассмотрим речевой корпус, разрабатываемый на филологическом факультете СПбГУ. *Один речевой день* (ОРД) — звуковой корпус современного русского языка повседневного общения. Корпус создается с целью изучения реаль-

ной речи носителей языка в естественных условиях коммуникации, и в этом состоит его отличие от абсолютного большинства речевых корпусов, записанных в лабораторных или в 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 [Асиновский, Богданова, Русакова и др., 2008]. Помимо речи информантов, в корпусе представлены записи их коммуникантов (родственников, друзей, коллег, знакомых и незнакомых), среди которых были люди самого разного возраста и разных специальностей. Общая длительность записанного материала — более 500 часов.

Данный корпус позволяет изучать лингвистическую динамику записанного материала: исследовать временные ряды количественных переменных с помощью стандартных статистических методов и анализировать частотные ряды (лексики, грамматических и, в частности, синтаксических структур, семантики или разговорных тем, тех или иных акустических явлений или просодических контуров) в зависимости от времени суток и условий коммуникации в самом широком понимании этого термина, а также решать множество других задач, таких как анализ влияния профессии на бытовую жизнь человека, получение информации о среднем артикуляционном темпе спонтанной речи носителей русского языка [Шерстинова, Рыко, Степанова, 2009].

В частности, при анализе корпусного материала можно обнаружить такие не зафиксированные в словарях новации, как междометная прагматема *щас* или вербальный хезитатив *щас-щас(-щас)* из редуцированной формы наречия *сейчас*; маркер-аппроксиматор *туда-сюда* также из соответствующего наречия (*он () он мне звонит / типа / короче / мы все готово / короче / туда-сюда / а я говорю / я в отпуске*); разнообразные прагматические значения местоимения-прилагательного *такой* (ксенопоказатель, вербальный хезитатив, изобразительный маркер, маркер-интенсификатор или деинтенсификатор) и мн. др. [Богданова-Бегларян, 2017].

Основной целью создания корпуса неподготовленных детских устных (извлеченных) текстов «Кондуйт» является получение базы устных неподготовленных текстов для изучения процессов формирования навыков построения связного текста в онтогенезе. Специфика корпуса «Кондуйт» (общий объем — 25,6 тыс. словоупотреблений) заключается в том, что в нем представлены образцы устной связной речи детей разных возрастных групп от двух до восьми лет. Всего в корпусе представлено 213 текстов, которые были получены в результате проведения серии экспериментов с детьми, посещающими дошкольные образовательные учреждения Санкт-Петербурга. Данный корпус можно использовать для проведения сравнительных исследований процесса формирования навыков организации связности и цельности текста, формирования разнообразия синтаксических структур, развития лексического запаса, формирования таких когнитивных способностей, как внимание к деталям, эмпатия и эмоциональная оценка поведения другого, развитие фантазийного мышления и многое другое, у русскоязычных детей в возрасте от двух до восьми лет [Эйсмонт, 2017].

#### 7.2.2.6. Мультимедийные 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. Кроме текстовой составляющей корпус такого рода может включать видео- и аудиозаписи процесса коммуникации с привязкой к тексту. Тексты выравнивают с их расшифровками, что позволяет исследовать не только языковые единицы, но и речевые действия говорящего в различных ситуациях общения, его неречевое поведение (мимику, жесты, позы).

В настоящее время большой интерес корпусных лингвистов привлекают способы передачи эмоций в устной речи — удивления, радости, огорчения и т. п. Примером корпуса, позволяющего проводить подобные исследования, является мультимедийный подкорпус в составе НКРЯ.

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 (морфологической, семантической, метатекстовой), стандартной разметки устных текстов (социологической, акцентологической) и стандартной разметки МУРКО (орфоэпической, разметки вокальной структуры), авторы провели разметку глубоко аннотированного МУРКО (разметку речевых актов, повторов, междометий и вокальных жестов, манеры говорения и др., разметку жестов) [Гришина, Савчук, 2008].

Мультимедийные корпусы являются перспективными с точки зрения исследования взаимодействия вербальной и невербальной составляющих 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 [Svartvik, Quirk, 1980], необходимо рассмотреть возможность 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 [Николаева, 2010]. Для целей исследования был создан **Корпус устных рассказов** на русском языке, стимулом для которых послужил шестиминутный видеосюжет «Фильм о груше» (“Pear film”). Об этом фильме было записано 8 рассказов студентов 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 [Крейдлин, 2002].

На примере из Корпуса устных рассказов исследователям удалось показать, как отдельные признаки жестов и положения рук могут давать дополнительную информацию об организации дискурса, состоянии говорящего и процессе 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 [Николаева, 2010].

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. п., позволяет изучить стратегии эмоционального взаимодействия и конфликта, непрерывное коммуникативное поведение, хезитации и речевые сбои. Его допустимо использовать как материал для обучения работников клиентских служб или как базу данных эмоциональных реакций для мультиплекаторов и режиссеров.

### ***7.3. 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. К специальным корпусам можно отнести и тематические, и жанровые, и видовые, различающиеся наполнением, которое связано с определенными задачами. Но существуют и сугубо «задачные» специальные корпусы.

Примером может быть **Санкт-Петербургский учебный корпус текстов школьников, изучающих английский язык (SPbEFLC)**, 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 с ноября по декабрь 2007 г. Авторами текстов являются 78 учеников 9–11-х 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 структурной бедности речевых произведений неносителей языка. В репертуаре грамматических структур, обнаруженных в SPbEFLC, есть такие, которые представляют собой случаи «переходной грамма-

тики» (интеръязыка), выражающиеся, например, в нарушении правил наполнения компонентов базовых структур. Так формируется ядро грамматики EFL (English as a Foreign Language), которое не совпадает с базовыми грамматическими структурами литературного английского языка. На основании корпусных данных авторы высказывают предположение о том, что складывающиеся нормы «глобального английского» во многом опираются на «окаменевшие» модели интеръязыка [Камшилова, 2010].

Сложным объектом с точки зрения создания и стандартизации являются исторические корпусы, такие как **Санкт-Петербургский корпус агиографических<sup>1</sup> текстов XV–XVII вв.** (СКАТ), доступный на сайте <http://ct05647.tmweb.ru/scat/page.php?page=project>. 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 [Азарова, Алексеева, Захарова, 2006].

В числе других специальных корпусов следует упомянуть Регенсбургский диахронический корпус русского языка (древнерусские тексты), корпус рукописных памятников Древней Руси (бестияные грамоты, летописи, рукописные книги), параллельный корпус переводов «Слова о полку Игореве», корпус русского электронного наследия «Манускрипт» [Баранов, 2019].

Также к специальным корпусам относятся корпусы, созданные на основе текстов определенного вида (жанра, типа) или определенной тематики. **Отраслевой** специальный корпус (кораблестроение, металлы, экология, навигация) дает специалисту самое главное — термины в их профессиональном окружении, контексты, из которых видно, что тот или иной автор имеет в виду под данным

---

<sup>1</sup> Агиография (от греч. ἅγιος «святой» и γράφω «пишу») — научная дисциплина, занимающаяся изучением житий святых, богословскими и историко-церковными аспектами святыни. Жития святых можно изучать с лингвистической, историко-богословской, исторической, социально-культурной и литературной точек зрения.

термином, какое понятие за ним стоит, а также статистические характеристики, что позволяет отследить изменения в терминологии, включая появление новых терминов.

### **Вопросы и задания для самоконтроля**

1. Дайте определения терминов: *лемматизация, парсинг, токен, treebank, WaC, wacky*.
2. Каковы основные процедуры обработки текстов на естественном языке при создании корпусов? Кратко охарактеризуйте каждую из них.
3. Перечислите названия бесплатных корпусных менеджеров.
4. Назовите причины, по которым веб может рассматриваться как огромный многоязычный корпус.
5. В чем суть технологии WaC?
6. Назовите два условия при обсуждении процедур обработки текстов корпусов.
7. Назовите и кратко охарактеризуйте наиболее известные корпусы русского и английского языков.
8. Опишите проект «Один речевой день». Какие экстралингвистические параметры текста значимы для корпусов устной речи?
9. Какие возможности предоставляет пользователям корпус МУРКО?
10. Какие характеристики Национального корпуса русского языка свидетельствуют о том, что этот корпус отвечает критерию репрезентативности и другим требованиям, предъявляемым к современным корпусам?

## Часть 3

# ПОЛЬЗОВАНИЕ КОРПУСАМИ

## Глава 8. Корпусные менеджеры

### 8.1. Корпус как поисковая система

Неотъемлемой частью понятия «корпус текстов» является **корпусный менеджер**, представляющий собой специализированную поисковую систему, включающую программные средства для поиска данных в корпусе, получения статистической информации и предоставления пользователю результатов в удобной форме.

В составе любой поисковой системы можно выделить три основные части:

1. Ввод документов — условное название подсистемы, обеспечивающей обработку и индексирование входного массива документов. В поисковых системах это еще и сканирование Интернета для формирования этого входного потока (этую функцию выполняет *робот-индексатор*).
2. Загрузчик — подсистема, обеспечивающая формирование *поисковой базы данных (индекса)* — специальным образом организованной структуры данных, включающей, прежде всего, инвертированный файл, состоящий из лексических единиц, взятых из проиндексированных документов, и содержащий разнообразную информацию об этих единицах (в частности, их позиции в документах), а также о самих документах и сайтах в целом.
3. Поисковый клиент — подсистема поиска, обеспечивающая обработку запроса пользователя (поискового предписания), поиск в базе данных и выдачу результатов поиска пользователю. Эта подсистема общается с пользователем через пользовательские интерфейсы — экранные формы, отобра-

жаемые в браузерах: выделяют интерфейс формирования запросов и интерфейс просмотра результатов поиска.

Обычно поисковой системой в обиходе называют именно этот третий компонент, который, по существу, не может функционировать без первых двух.

Все эти компоненты существуют и в корпусных системах. Ввод документов (текстов) для корпуса по традиционной технологии включает процедуры 1–7 (см. п. 5.1). Согласно технологии WaC (см. п. 6.3) им соответствуют процедуры 1–9. Затем обработанные таким образом тексты поступают на вход «Загрузчика» и загружаются в поисковую базу данных, соответственно, это гл. 9 (разд. 5.1) и гл. 10 (разд. 6.3). Далее в этой базе данных (обычно называемой просто корпусом) осуществляется поиск с помощью корпусного менеджера. Причем следует отметить, что тексты, обработанные и размеченные в подсистеме ввода, могут быть загружены в разные корпусные менеджеры.

Корпусный менеджер как поисковая система в широком смысле слова — это программно-лингвистические средства, обеспечивающие в целом создание и использование корпусов. Эта система обработки и управления корпусными данными представляет собой набор разнообразных программных средств, которые способны работать автономно или могут быть объединены в один или несколько комплексов. Функции их разнообразны. На самом верхнем уровне можно выделить следующий набор функций:

1. предварительная обработка текстов;
2. автоматическая разметка;
3. загрузка данных в корпус;
4. поиск в корпусе и выдача результатов;
5. сервисные функции.

Корпусный менеджер в узком смысле слова — это специализированная система, реализующая функции типа 4 и частично 5.

Таким образом, говоря об инструментарии корпусной лингвистики, необходимо либо иметь в виду единый комплекс, который реализует все функции, что бывает крайне редко, либо разработать (адаптировать) и совместить соответствующие программные средства для каждой процедуры.

Если рассмотреть известные сегодня корпусные менеджеры под углом зрения этих процедур, то мы увидим, что некоторые про-

граммы интегрируют несколько из вышеперечисленных функций, а некоторые строго однофункциональны. Так, корпусный менеджер DDC (<http://aot.ru/download/concord.zip>) выполняет функции 2, 3, 4. Система Sketch Engine фактически реализует все функции. Ее загрузчик не только проводит морфологическую разметку и формирует собственно корпус, но и выполняет работу по преобразованию файлов на входе, например их разархивацию.

Основной результат поиска в корпусе — это конкорданс. Однако кроме этого доступно получение справок о характеристиках текста или лексических единиц, статистических данных о языковых единицах и о лингвистических категориях и метаданных (частоте словоформ, лексем, грамматических категорий, изменениях частот и контекстов в различные периоды времени, данных о совместной встречаемости лексических единиц, жанрово-стилистических характеристиках и т. п.). Эти статистические данные могут выдаваться непосредственно (например, частотный список) или использоваться для «внутренних» подсчетов и выдачи новых данных: количественных характеристик устойчивости сочетаний в тексте, парадигматических семантических кластеров лексических единиц, ключевых слов исследуемого корпуса и т. п., непосредственно в корпусе не заложенных.

Корпусный менеджер в узком смысле слова должен:

- строить конкорданс (список контекстов);
- искать контексты не только по отдельным словам, но и по словосочетаниям;
- давать возможность отображать найденные словоформы в широком контексте;
- осуществлять поиск по сложным запросам, включающим лексические единицы, грамmemы, логические операторы;
- сортировать полученные списки по нескольким критериям, выбранным пользователем;
- давать статистическую информацию по отдельным элементам корпуса и по всему конкордансу;
- 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;

- быстро обрабатывать запросы и выдавать результаты;
- быть легким (интуитивно понятным) в использовании (это актуально как для опытного, так и для начинающего пользователя).

Наиболее известны такие универсальные корпусные менеджеры, как Sketch Engine (старое название — Manatee/Bonito), No-Sketch Engine, CQP, AntConc, MonoConc, ParaConc, WordSmith Tools. Под универсальностью мы здесь понимаем возможность использования менеджера для создания любых корпусов и степень его распространенности. Менее известен корпусный менеджер системы Corpus Technologies, на базе которого тем не менее создано значительное число корпусов малых языков. Менеджеры могут разрабатываться как специально, так и на основе систем управления базами данных (СУБД) или поисковых систем. Например, поиск в Национальном корпусе русского языка до последнего времени осуществлялся поисковой системой Yandex.Server 3.8 Professional [Национальный корпус русского языка]. С сентября 2019 г. НКРЯ перешел на новую технологию поиска.

Т. Мак-Энери и А. Харди [McEnery, Hardie, 2012] описывают четыре поколения корпусных программных средств. Большинство современных инструментов, используемых корпусной лингвистикой, классифицируются как инструменты третьего поколения. Они предлагают множество функций, включая статистические методы, обладают некоторой масштабируемостью для работы с большими корпусами, предлагают многоязычную поддержку и дружественный интерфейс. Примеры таких инструментов — WordSmith Tools [Scott, 1996], MonoConc Pro [Barlow, 2000] и AntConc [Anthony, 2019]. Наибольшим ограничением является то, что они плохо работают с большими корпусами, и если раньше, в корпусах второго и отчасти третьего поколения, можно было говорить о решении корпусных задач даже такими «подручными» средствами, как команды обработки текста *grep*, *sort*, *unix* или скрипты на Питоне, то для современных колossalных корпусов требуется совсем другая системная архитектура.

Сегодня больших и сверхбольших корпусов становится все больше ввиду потребности в них, увеличения объема электронных документов и развития технологий типа *wacky*. Ответом на это стало создание инструментов четвертого поколения, таких

как corpus.byu.edu, CQPweb [Hardie, 2012], Sketch Engine [Kilgariff, 2013], Wmatrix [Rayson, 2013]. Эти инструменты предлагают лучшую масштабируемость за счет хранения корпуса в базе данных веб-сервера и предварительной индексации данных для обеспечения быстрого поиска.

Несмотря на перечисленные выше преимущества, инструменты четвертого поколения имеют ряд ограничений. Если пользователь хочет составить свой небольшой корпус и выполнить на нем простой анализ, то он лишен этой возможности. По крайней мере, для доступа к серверу ему придется зарегистрироваться на сервисе, подписать различные лицензии и, возможно, перечислить ежемесячную абонентскую плату. Альтернативой является установка инструмента на персональный сервер, но для этого нужно получить (купить) корпусный менеджер, сервер, настроить их, установить программное обеспечение и поддерживать работоспособность сервера.

Еще одна проблема заключается в том, что в инструментах четвертого поколения размыты границы между данными и инструментом. Из-за способа хранения данных в индексированной форме на внешнем сервере пользователи не имеют возможности обратиться непосредственно к исходным данным, по крайней мере быстро соотнести их с результатами поиска в корпусе.

Заслуживают внимания результаты опроса лингвистов с просьбой указать, какие корпусные компьютерные средства они чаще всего используют для научных исследований [Tribble, 2008] (рис. 8.1).

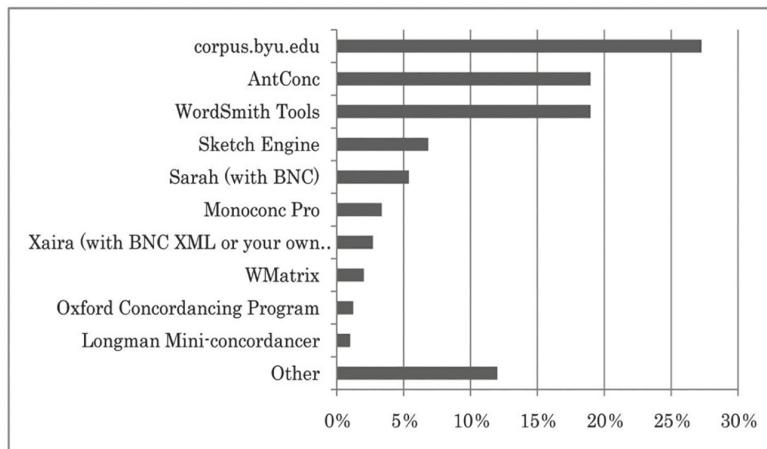


Рис. 8.1. Наиболее популярные средства для работы с корпусами

Вероятно, данные рис. 8.1 требуют уточнения путем проведения нового опроса. При этом следует заметить, что наиболее популярный [сорпус.bsu.edu](http://corpus.bsu.edu) не является открытым корпусным менеджером, а предоставляет корпусную службу с множеством корпусов (в первую очередь английского языка) с закрытой системой управления корпусными данными, то есть его популярность во многом определяется большим числом исследователей, интересующихся английским языком, а также его функциональными возможностями.

## **8.2. Функциональные возможности корпусных менеджеров**

Покажем характерный набор функций корпусных менеджеров, относящихся в основном к четвертому поколению, который является обобщением анализа разных систем. Понятно, что выделить их не так просто. Еще сложнее привести их к одному знаменателю. Сложности очевидны: одни и те же функции в разных корпусных менеджерах могут называться по-разному. В разных корпусных менеджерах эти функции могут решаться в разных объемах и на выходе выдавать похожие, но не идентичные результаты. Следует также понимать, что если в одной системе некоторая функция имеется и реализуется на полном (любом) объеме корпуса, то наличие такой же функции в другой системе позволяет ее выполнить лишь на корпусе ограниченного объема.

Итак, вот как примерно выглядит функционал современных корпусных менеджеров.

### **Функции**

#### ***Создание корпусов***

#### ***Поиск по одному или нескольким корпусам:***

- поиск по словоформе;
- поиск по лемме;
- поиск по нескольким словам;
- поиск по символам (по последовательности символов);
- поиск по аффиксам;
- поиск с помощью регулярных выражений;
- поиск в параллельных корпусах.

***Формирование и обработка конкорданса:***

- возможность задать величину контекста (по символам и словам);
- просмотр расширенного контекста для заданного слова;
- сортировка;
- график конкорданса;
- просмотр текстового файла;
- информация о частоте заданного слова.

***Выделение коллокаций:***

- поиск с указанием части речи коллоката;
- поиск для разной ширины контекстов;
- ранжирование по мерам ассоциации.

***Сравнение слов в корпусах:***

- сравнение по частоте;
- сравнение по грамматическим характеристикам.

***Другие функции:***

- создание частотного списка слов;
- лексико-синтаксические шаблоны;
- кластеризация;
- выделение N-грамм;
- выделение ключевых слов;
- классификация текстов по жанрам;
- выделение слов, сходных по значению (слов одного лексико-семантического поля).

### **8.3. Языки запросов корпусных менеджеров**

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.

Основной результат поиска по запросу в корпусе — это конкорданс и его объемная характеристика (количество строк и относительная частота искомого элемента — *ipm*).

Согласно словарю иностранных слов, *конкорданс* — это расположенный в алфавитном порядке перечень встречающихся в книге слов с минимальным контекстом (в несколько слов). Конкорданс в такой форме обычно называется KWIC (Key Words in Context). В словаре Collins Cobuild English Dictionary слово *concordance* определяется следующим образом: алфавитный список слов в книге или комплекте книг, в котором указано, где каждое слово находится и как часто оно используется.

В корпусной лингвистике *конкорданс* — это список всех вхождений в корпусе (контекстов) заданного в запросе языкового выражения (слово, словосочетание, запрос в виде сложной формулы), возможно, со ссылками на источник.

Обработку полученного конкорданса можно также отнести к основным функциям корпусных менеджеров. Покажем характерный набор функций корпусного менеджера на примере трех систем: AntConc (<https://www.laurenceanthony.net/software/antconc/>), Intellitext (<http://corpus.leeds.ac.uk/it/>) и Sketch Engine (табл. 8.1).

Таблица 8.1. Поисковые возможности корпусных менеджеров

Функция	Корпусный менеджер		
	AntConc	IntelliText	Sketch Engine
<i>Поиск по одному или нескольким корпусам</i>			
Поиск по словоформе	+	+	+
Поиск по лемме	-	+	+
Поиск по нескольким словам	+	+	+
Поиск по символам (последовательности символов)	+	-	+
Поиск по аффиксам	-	+	-
Поиск с помощью регулярных выражений	+	+	+
Поиск в параллельных корпусах	-	-	+

Окончание табл. 8.1

Функция	Корпусный менеджер		
	AntConc	IntelliText	Sketch Engine
Возможность задать величину контекста (по символам и словам)	+	+	+
Просмотр расширенного контекста для заданного слова	+	+	+
Сортировка	+	+	+
График конкорданса	+	-	-
Просмотр текстового файла	+	-	-
Информация о частоте заданного слова	+	+	+

Можно сказать, что левый столбец фактически представляет собой обобщенный сводный перечень поисковых функций современных корпусных менеджеров. Покажем эти функции на примере языка запросов системы Sketch Engine.

#### 8.4. Язык запросов корпусного менеджера Sketch Engine

Корпусный менеджер Sketch Engine представляет собой программное средство для работы с корпусами текстов. Он состоит из различных модулей, в данном разделе рассматривается только функция поиска. Для демонстрации работы с системой используется корпус русских текстов по корпусной лингвистике, созданный М. В. Хохловой на кафедре математической лингвистики СПбГУ.

Поисковый интерфейс системы Sketch Engine показан на рис. 8.2. С 20 января 2020 г. Sketch Engine перестает поддерживать этот интерфейс и полностью переходит на новый (см. <https://auth.sketchengine.eu/>). Этот интерфейс отображается после того, как пользователь выбрал корпус.

Интерфейс поиска в корпусе состоит из верхней панели, стандартной для всех режимов (основная функция — выбор корпуса из списка ранее использованных и указание на открытый в данный момент корпус), левого меню и основного поискового окна. Левое

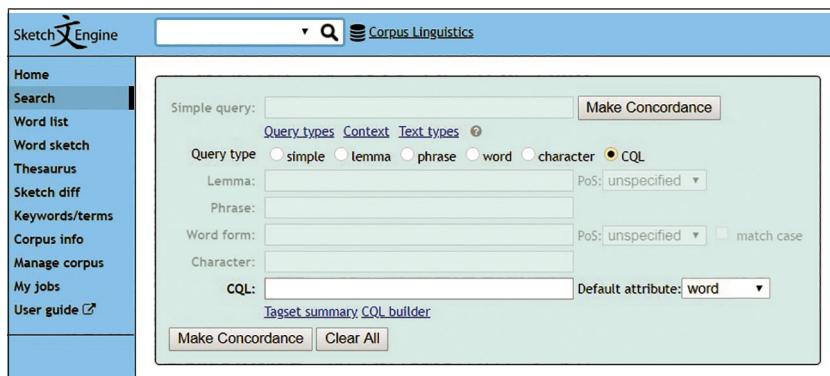


Рис. 8.2. Поисковый интерфейс системы Sketch Engine

меню обеспечивает переключение между доступными функциями, список функций зависит от выбранного корпуса и от функции, выполненной на последнем шаге. В основном поисковом окне необходимо ввести поисковый запрос и выбрать тип запроса (Query type): основной (*basic*), поиск по лемме (*lemma*), по словосочетанию (*phrase*), по словоформе (*word*), по символу или сочетанию символов (*character*) или на языке CQL.

Поиск по лемме выдает контексты, в которых заданное слово встретилось в любой форме, по словосочетанию и по словоформе — именно в той форме, в какой словосочетание или словоформа заданы. Тип запроса «основной» обеспечивает поиск как по лемме или по словоформе, так и по сочетанию лемм или словоформ в зависимости от того, что введено в окно запроса, при этом регистр значения не имеет.

Кроме того, имеются дополнительные режимы «Контекст» (Context) и «Тип текста» (Text types) (рис. 8.2). «Контекст» (рис. 8.3) позволяет задать дополнительные условия поиска в виде ограничений на контекст в виде одной или нескольких лемм либо в виде одной или нескольких частей речи, причем это ограничение действует в пределах заданного контекстного окна (*window*) (левого, правого или двустороннего). При этом сами леммы или части речи могут объединяться операциями конъюнкции (*all*), дизъюнкции (*any*) или отрицания (*none*).

Документы в корпусе могут быть размечены метаданными (жанр, время публикации, автор, имя исходного файла и т. д.), и тог-

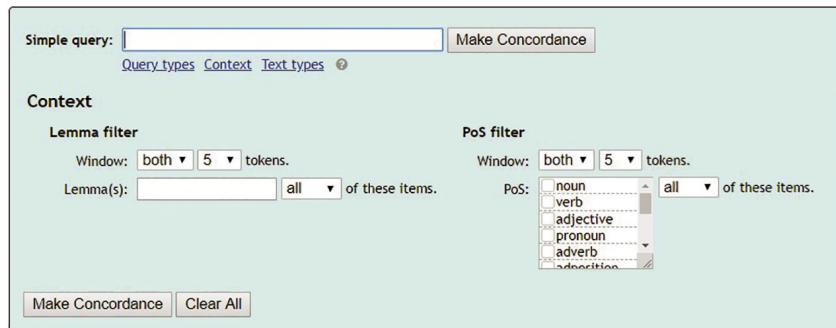


Рис. 8.3. Поисковый интерфейс системы Sketch Engine (ограничения по контексту)

да появляется возможность ограничить поиск данными метаразметки (пункт меню Text types). Внутри метаданных одного и того же типа предполагается операция дизъюнкции, между разными типами — операция конъюнкции.

Язык CQL (Corpus Query Language) — это язык, позволяющий сформулировать сложный запрос на поиск в терминах лемм, словоформ, морфологических тегов, полных или усеченных, и любых других атрибутов, присутствующих в разметке корпуса [Jakubíček, Kilgarriff, McCarthy et al., 2010]. В любом месте запроса могут быть применены логические операторы конъюнкции, дизъюнкции и отрицания (И/ИЛИ/НЕ). Язык запросов позволяет также описывать разрывные сочетания, задавать ограничения на контекст и многое другое. Подробное описание языка запросов CQL и языка регулярных выражений см. на сайте системы (<https://www.sketchengine.eu/documentation/corpus-querying/>). Язык CQL в Sketch Engine, равно как и режимы фильтрации при формировании частотных списков (*word list options*), реализован на основе языка регулярных выражений (см. п. 8.5). На сайте системы можно также найти упражнения по использованию этого языка (<http://regex.sketchengine.eu/>).

Покажем пример запроса на языке CQL:

```
[tag="V.*"] [word!="[:punct:]"]&tag!="V.*" {0,3}  
[lemma="для"] [tag="A.*|P.*"]{0,2} [tag="N.*"]
```

Этот запрос задает системе следующие условия поиска: найти конструкции с предлогом «для», которым управляет глагол в лю-

бой форме (`tag="V. *"`) и которому подчинено существительное в любой форме (`tag="N. *"`), причем между глаголом и предлогом может быть не более трех токенов (каждая позиция обозначена квадратными скобками), но этими токенами не могут быть (!= ‘не равно’) знак препинания (`[[:punct:]]`) и другой глагол, а между предлогом и существительным должно быть не более двух слов, и этими словами могут быть только прилагательные и/или местоимения (`tag="A. * | P. *"`). И этот пример демонстрирует лишь незначительную часть возможностей языка CQL.

Для кодировки грамматических признаков для большинства языков используется стандарт MULTEXT-East morphosyntactic specifications 4.0 (<http://corpus.leedsac.uk/mocky/msd-ru.html>).

В результате поиска получается конкорданс, в нашем случае объемом в 521 строку (рис. 8.4, см. с. 122).

Конкорданс — это список всех примеров запрашиваемого слова или конструкции, найденных в корпусе и сопровождаемых некоторым (настраиваемым) контекстом слева и справа. На приведенном рисунке показан наиболее часто используемый формат KWIC.

Над этим конкордансом могут быть выполнены разнообразные операции, указанные в нижней половине левого меню (рис. 8.5, см. с. 123), например, случайное перемешивание строк конкорданса (*shuffle*), какие-либо дополнительные ограничения на него (*filter*), подсчет частот полученных лексических конструкций (*frequency*) (рис. 8.6, см. с. 123), подсчет частоты левых и правых окружений, вычисление силы синтагматической связанности между лексемами (*collocations*).

Подробнее о возможностях и способах работы с системой Sketch Engine, в том числе и о формулировании запросов, см. в работе Дж. Томаса [Thomas, 2016].

## 8.5. Язык регулярных выражений RegEx

Развитые языки запросов корпусных менеджеров, как правило, базируются на формализме, который получил название «язык регулярных выражений» (RegEx) [Фридл, 2001; Смит, 2006].

RegEx — это формальный язык поиска и осуществления манипуляций со строками, основанный на использовании метасимволов (*wildcard characters*). Для поиска используется строка-образец («шаблон», «маска», англ. *pattern*), состоящая из символов и метасимво-

Query V.\*, [[:punkt;]], Для, А.\*|Р.\*|N.\* 521 > Shuffle 521 > Shuffle 521 > Shuffle 521 > Shuffle 521 (1.516.99 рег million) ①

	Page 1	of 27	Go	Next   Last
file#4	Д.Хармса в формате TEI	использован для литературоведческой разметки	комментарiev и Т.Н. «мотивов	
file#5	разметку планируется	использовать для анализа	контекстов предикатных	
file#0	. Различные корпусы могут	использоваться для получения	разнообразных	
file#0	на то, что это предположение	представляется слишком сильным для двуязычного корпуса	текстов по одной и той же	
file#1	быть легко собраны вместе и	стать одновременно доступными для пользователей	Такую совокупность	
file#3	непосредственно теми, кто	занимаются подготовкой текстов для корпуса	и не выносится на открытый	
file#3	• Subgenre: точнее	определяет жанр текста для жанров	поу, col, ess:	
file#5	друг с другом. Корпус может	использоваться как материал для обучения	работников клиентских служб	
file#5	], не обавательно целью, то	есть для согласования	экспериментальных	
file#3	существующие корпусы мало	подходит для нужд	диахронической	
file#3	в одном направлении: корпусы	создаются для нужд	фундаментальной	
file#3	бы, не очень важно, какой язык	применяется для кодирования	промежуточных данных,	
file#2	методов обработки текста,	имеют большое значение для анализа	дискурсивной системы.	
file#5	ELAN Все рассказы корпуса	будут доступны для скачивания	в формате ELAN - одном из	
file#2	• Вместо этого, они	служат только для маркировки	аспектуального значения (	
file#3	категория. Категории, не	являющими существенными для указанной части	речи, обозначаются знаком «-	
file#5	, beyond, over, также могут	использоваться для обозначения	увеличения: go beyond '	
file#3	набор параметров	служит достаточным основанием для построения	типов генетиков и	
file#0	... Поэтому латинский язык	использовал для образования	форм претерита инфекта	
file#3	затруднительным, была	разработана система тегов для разметки	текстов, поступающих в АЛК.	

Page 1 of 27 Go Next | Last

Рис. 8.4. Конкорданс KWIC для вышеприведенного запроса

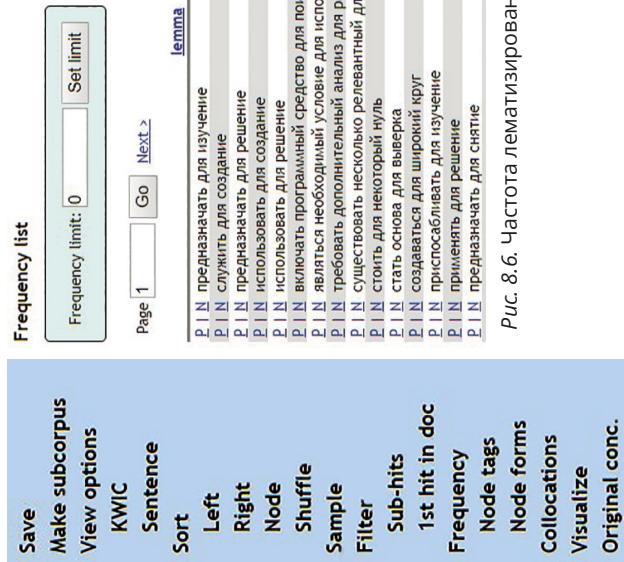


Рис. 8.6. Частота лемматизированных конструкций

Рис. 8.5. Функции  
системы  
Sketch Engine,  
выполняемые над  
конкордансом

волов и задающая правило поиска (Википедия, [https://ru.wikipedia.org/wiki/Регулярные\\_выражения](https://ru.wikipedia.org/wiki/Регулярные_выражения)). Причем нужно иметь в виду, что существуют разновидности регулярных выражений и нередко реализации RegEx привязаны к языкам программирования. При этом большую часть запросов на языке RegEx разработчики «скрывают» от пользователя в программном коде, реализовав их в виде удобного интерфейса. Пользователю необходимо лишь заполнить определенные поля формы (веб-страница с формами для заполнения), и его запрос будет осуществлен. Но для задания сложных запросов, в том числе на языках запросов корпусных менеджеров, полезно знать основы языка регулярных выражений.

**Регулярные выражения** — это строковые записи, задающие правила поиска. 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* (*matching*) 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. RegEx реализован в самых разных языках программирования и СУБД. 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:

- *простые символы*, а также *управляющие символы*, играющие роль их заменителей;
- *управляющие конструкции* (квантификаторы повторений, оператор альтернативы, группирующие скобки и т. д.);
- 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. Любой символ в строке на языке RegEx обозначает сам себя, если он не является специальным (см.: <https://ru.wikipedia.org/>

wiki/Регулярные\_выражения «Синтаксис. Представление символов»). Например, по регулярному выражению “abcd” будут найдены совпадения во всех строках, 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”.

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”, или “а” и, наконец, “к”. Если буквальтернатив 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” в нижнем регистре, а выражение “[а-я0-9]” добавляет к ним любую арабскую цифру.

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:

- \d — любой цифровой символ;
- \D — нецифровой символ;
- \s — пробел;
- \S — любой непробельный символ;
- \w — буквенный или цифровой символ либо знак подчеркивания;
- \W — любой символ, кроме буквенного или цифрового символа или знака подчеркивания.

**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 определения того, сколько раз в строке могут встречаться предшествующие им символы первого класса.

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

*Одно совпадение и более.* Символ плюса “+” обозначает одно совпадение или более 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[а]?” будут соответствовать строки “Петров”, “Петрова” и “Петровы”.

*Заданное число совпадений.* Последний квантификатор повторения — фигурные скобки “{}”. С его помощью можно задавать правила квантификации более точно. 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  $n$  раз или более;

$A\{,m\}$  — символ “A” может быть повторен не более  $m$  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 “|”, например “давать|давал|давала|давали|давали” соответствует всем подстрокам, разделенным символом альтернативы “|”, а выражение “1|2|3” соответствует подстрокам 1, 2 или 3, что эквивалентно выражению [123].

**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”, выражение “^перенять\$” — точному совпадению со строкой “перенять”, а выражение “^пре|^пере” — строкам, которые начинаются с “пре” или с “пере”.

Более подробно о регулярных выражениях, кроме вышеуказанной литературы, см.: <https://www.regular-expressions.info/index.html> и <http://nikic.github.io/2012/06/15/The-true-power-of-regular-expressions.html> (русский перевод: <https://habr.com/ru/post/171667/>).

## 8.6. Сервисные функции

Сервисные функции и функциональность корпусных служб во многом базируются на обработке статистических данных. Естественно, даже на современных мощных компьютерах обработка корпусных данных, насчитывающих сотни миллионов и даже миллиарды словоупотреблений, — это нетривиальная задача. Однако еще большую сложность представляет «интеллектуализация» работы систем, а именно выявление лингвистических, по сути семантических, закономерностей того или иного языка.

Статистику в корпусных службах условно можно поделить на внешнюю и внутреннюю. **Внешняя статистика** — это то, что выдается непосредственно пользователю как отчет, как справка, как та или иная количественная характеристика корпуса и языка (подъязыка), который корпус представляет. Эти функции, например получение статистической информации о корпусах или подкорпусах, могут выполняться по запросу пользователя (например, Sketch Engine) или представлять собой справочный раздел на сайте (например, Национальный корпус русского языка).

С точки зрения внешней статистики, одну из наиболее полных статистических картин распределения текстов по элементам метаданных дает НКРЯ, предоставляющий сведения о распределении тек-

стов по подкорпусам, по видам текстов, по жанрам, по сфере функционирования, по тематике, по хронологическим периодам. Все эти распределения можно посмотреть в терминах числа текстов, предложений, числа и процентной доли словоупотреблений по данной характеристистике по отношению к общему объему. Эти характеристики НКРЯ суммарно выложены непосредственно на сайте, но частично их можно получить и по результатам отдельного поиска.

Второй тип статистических данных — это частотные списки слов или *n*-грамм. Здесь следует выделить систему Sketch Engine. Выдавая характеристику «объем корпуса», она указывает как количество токенов, так и количество слов, то есть чисто лексических токенов. Она позволяет также получить самые разные частотные списки по выбранному корпусу: частотный список словоформ, лемм, и то и другое без учета регистра или с учетом (в этом случае одно и то же слово, начинающееся с прописной или строчной буквы, считается разными словами), а также частотный список всех тегов (токенов с одинаковыми тегами).

Почти все исследованные корпусные инструменты умеют составлять частотные списки *n*-грамм. Это или списки, составленные по всему корпусу и предоставляемые как архивы с текстовыми файлами (НКРЯ, система Марка Дэвиса Corpus.buu.edu), или порождаемые динамически по корпусу или подкорпусу (Sketch Engine, AntConc).

**Внутренняя статистика** используется для получения собственно лингвистических результатов, таких как кластеризация лексических единиц, выявление коллокаций, ключевых слов и т. п. Внутренняя статистика предоставляет данные, которые используются для «внутренних» подсчетов и выдачи новых данных, непосредственно в корпусе не заложенных.

Покажем эти сервисные функции на примере системы Sketch Engine, наиболее продвинутой в этом отношении. Помимо стандартной функции генерации конкорданса и представления его в разных видах (Concordance Search), генерации частотных списков лемм, словоформ, *n*-грамм, тегов (Word List), Sketch Engine включает следующие инструменты:

- Collocations выявляет статистически устойчивые словосочетания, связанные с термином запроса с использованием различных мер ассоциации.

- Word Sketch генерирует списки коллокаций в разрезе заданных синтаксических конструкций.
- Word Sketch Difference (Sketch diff) показывает различия в сочетаемости заданной пары слов.
- Thesaurus генерирует список слов, более всего семантически связанных с заданным (лексико-семантическое поле). Этот инструмент включает в себя также «Кластеризацию» (Clustering) — группировку единиц тезауруса в кластеры — лексико-семантические группы.
- Keywords and Terms выявляет ключевую лексику с использованием *keyness score*.
- Trends анализирует использование слов в диахронии (для соответствующим образом размеченных корпусов) [Herman, Kovář, 2013].
- WebBootCaT составляет комплекс программ для создания пользовательских корпусов на основе текстов из веба [Kilgarriff Baisa, Bušta, 2014a; Kilgarriff et al., 2014b].

Все инструменты выдают результаты своей работы с количественными параметрами.

Подробное описание работы с этой системой можно найти в книге Дж. Томаса [Thomas, 2016]. Эта книга представляет собой практическое введение в корпусную лингвистику, построенное вокруг Sketch Engine. Автор не только показывает корпусные инструменты, но и побуждает читателя к исследованию. Обширный список литературы делает этот учебник теоретически обоснованным. Обширный иллюстративный материал (снимки экрана) и разнообразные задания делают его практически полезным.

На рис. 8.7 приведен пример автоматического построения дистрибутивного тезауруса по корпусу текстов «Понятие империи в русской культуре» [Захаров, 2018] для лексемы «империя» с автоматическим разделением его на кластеры. Для вычисления парадигматического подобия слов рассматриваются наборы сочетаемости для пар слов с учетом синтаксического отношения (лексические шаблоны). Единицы семантического поля обладают общими синтагматическими и парадигматическими свойствами, что отражает их семантическую близость. Схожесть дистрибуции слов высчитывается статистически на основе меры ассоциации logDice [Rychlý, 2008] и с учетом лексико-синтаксических шаблонов [Kilgarriff, Rychlý, 2007].

империя			
Rise_new_2 freq = 989			
Lemma	Score	Freq	Cluster
государство	0.31	1182	страна [0.203, 807] европа [0.172, 921] религия [0.161, 573] общество [0.139, 697] человечество [0.121, 427]
культура	0.213	681	цивилизация [0.207, 460] литература [0.155, 422] просвещение [0.119, 317] философия [0.114, 254] наука [0.11, 395]
мир	0.202	1596	церковь [0.183, 2088] народ [0.174, 2456] племя [0.131, 835] человек [0.117, 2388]
рим	0.197	662	русь [0.186, 704] византия [0.122, 323]
россия	0.186	2632	
монархия	0.178	211	христианство [0.151, 495] православие [0.108, 435]
история	0.177	1447	жизнь [0.155, 1999] развитие [0.127, 805]
царство	0.176	492	
власть	0.165	1168	
революция	0.151	388	война [0.12, 449]
император	0.141	272	царь [0.124, 525]
идея	0.134	931	политика [0.114, 238] мысль [0.109, 961]
город	0.134	449	раскол [0.116, 172]
семья	0.133	361	нация [0.128, 238]
учение	0.133	545	дух [0.12, 909] вера [0.116, 1025]

Рис. 8.7. Гнездо тезауруса с выделенными кластерами для ключевого слова *империя*

В первом столбце приведены лексемы, во втором — значение статистической меры logDice, в третьем — абсолютная частота лексемы в корпусе, в четвертом — лексемы, образующие с лексемой из первого столбца единый кластер (в квадратных скобках указаны значение меры и частоты, им соответствующие). Можно заметить, что внутри кластеров сила и природа парадигматических связей разная: встречаются слова-синонимы, которые взаимозаменяемы в ряде контекстов, и слова, связанные другими отношениями, например гипонимией или меронимией. Примером первого типа могут служить лексемы *император* и *царь*, *государство* и *страна*.

Однако и те слова, которые мы относим к синонимам, не являются абсолютными синонимами с точки зрения употребления их в речи. Инструмент «Дифференциация» (Sketch diff) позволяет выявить сходство и различие в сочетаемости для пар слов, в том числе для тех, которые мы считаем синонимами. Покажем это на примере пары *государство* и *страна* (рис. 8.8).

Мы видим, что практически во всех синтаксических отношениях эти слова показывают разную сочетаемость. Так, в отношении *gen\_modifies* (исследуемое слово стоит в родительном падеже) если *государство* сочетается с такими словами, как *глава государства*, *падение*, *нужда*, *интерес*, *политика* и т. д., то для *страны* характерны сочетания со словами *центр* (*центр страны*), *население*,



Рис. 8.8. Различия в сочетаемости для слов государство и страна

ряд, житель, капитализация и только слова история, судьба, развитие сочетаются как со страной, так и с государством.

Итак, можно констатировать, что инструментарий современной корпусной лингвистики обладает большими функциональными возможностями и позволяет получать разнообразные сведения, характеризующие и норму, и узус языка. Прогресс в сфере компьютерных технологий влечет за собой прогресс в создании и совершенствовании средств автоматической обработки текста и, как результат, порождает новые парадигмы лингвистических исследований.

Примеры использования статистических данных для решения разных лингвистических задач см. в п. 13.3.

## Глава 9. Способы использования корпусов

### 9.1. Пользователи корпусов

Пользователей корпусов, в первую очередь лингвистов, 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, проверка орфографии и грамматики и т. п. Так, статистическими методами на материале корпуса можно определить, какие слова регулярно встречаются вместе и поэтому могут быть отнесены к устойчивым словосочетаниям. Устойчивые словосочетания представляют собой неделимую с семантической точки зрения 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 лексикологии и семантики. Наблюдая окружение той или иной лингвистической единицы в корпусе, можно установить определенные признаки (маркеры), характеризующие семантику данной единицы.

Лингвисты-теоретики используют корпусы в качестве экспериментальной базы для проверки гипотез и доказательства своих теорий. Прикладные лингвисты (преподаватели, переводчики) 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 текстов 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 текстов, как период, автор или жанр. Литературоведы используют корпусы для стилеметрических исследований. Наконец,

корпусы используются в компьютерной лингвистике и информационных технологиях для разработки и настройки различных автоматизированных систем (машинного перевода, распознавания речи, информационного поиска).

Корпусы не могут заменить самонаблюдение (интроспекцию) ученого и обеспечить суждения лингвистов о лексике и грамматике, но они дают специалистам богатый репрезентативный эмпирический материал. Корпусы дают три типа данных, которые используются в ходе лингвистических исследований: эмпирическую поддержку, информацию о частотности, экстралингвистическую информацию (метаинформацию). Рассмотрим эти типы данных более подробно.

## 9.2. Что можно получить из корпуса?

### 9.2.1. Эмпирическая поддержка

Многие лингвисты используют корпус как банк примеров, то есть 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 на основе корпусных данных. Так, в корпусах текстов было найдено достаточно много грамматически правильных примеров начальной позиции частицы [Lüdeling, Kytö, 2008]. Подобно этому, предложения, объявленные сторонниками генеративной лингвистики грамматически неправильными, скорее должны считаться грамматически правильными, потому что подобные структуры на самом деле регулярно встречаются в современном английском языке [Карлсон, 2009]:

- Harry reminds me of himself [Postal, 1970]; cp.: Joe reminds me of himself (Интернет).
- John will leave until tomorrow [Lakoff, 1970, c. 148]; cp.: I will leave until tomorrow (Интернет).

- What an idiot I thought Tom was [Postal, 1968, с. 75]; cp.: What an idiot I thought the main character to be (Интернет).

Об этом же пишет Н. В. Перцов, опровергающий суждения авторитетнейших лингвистов о русском языке, используя материал Национального корпуса русского языка по состоянию на начало 2006 г.: «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... Очень часто встречающиеся в такого рода публикациях утверждения о фактах языка противоречат корпусным данным» [Перцов, 2006, с. 318–319]. И далее автор дает подборку случаев расхождения суждений о русском языке, высказанных нашими крупнейшими лингвистами, с данными, извлеченными из Национального корпуса русского языка. Утверждения о фактах языка, расходящиеся с данными корпуса, относятся к разным уровням языка. Здесь можно встретить утверждения о лексической сочетаемости или об управлении конкретных лексем, о значении слов, о семантической сочетаемости, об особенностях синтаксических конструкций, о грамматическом словоизменительном значении. Вот только некоторые из них:

- **нельзя** длинные глаза, **нужно** продолговатые глаза [Труб, 2006, с. 71]: cp. НКРЯ: Глеб вздрогнул: его **длинные глаза** какое-то время словно проверяли что-то во мне, ранее подвергавшееся сомнению (Е. Маркова, 2000)<sup>1</sup>. При этом в корпусе было найдено 14 якобы неприемлемых контекстов и 6 — с «нормальным» словосочетанием.
- **нельзя** тонкие колени, **нужно** острые колени [Труб, 2006, с. 71]: cp. 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» [Зализняк, Левонтина, Шмелев, 2005, с. 164]:

---

<sup>1</sup> В скобках приводится автор произведения, в тексте которого найдены искомая единица или фрагмент, и год издания.

ср. НКРЯ: *...в России счастье, по прогнозам российского президента, наступит только в 2010 году* («Известия», 30.10.2001).

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.

### 9.2.2. Статистическая информация

Эмпирическая поддержка представляет собой качественный метод использования корпуса, но корпусы также подкрепляют ее информацией о частотности для слов, фраз и конструкций, которая может быть использована для разнообразных исследований. Количественные исследования (которые, конечно, сочетаются с качественным анализом) используются во многих сферах теоретической и компьютерной лингвистики. Они показывают сходство и различие между разными группами говорящих или между разными типами текстов, обеспечивают данные о частотности лексических единиц и конструкций для психолингвистических исследований.

### 9.2.3. Метаинформация

В дополнение к лингвистическому контексту корпус представляет экстралингвистическую информацию, или метаинформацию, по таким факторам, как возраст или пол говорящего/пишущего, жанр текста, временная или пространственная информация о происхождении текста и т. д. Она позволяет сравнивать разные типы текстов или разные группы говорящих, а также просматривать и анализировать полученные данные (конкорданс) с учетом метаданных, относящихся к каждому отдельному употреблению.

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 использу-

ется в области психолингвистики учеными, которые интересуются тем, как дети овладевают языком [McWinney, 2000].

*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 [Johansson, 2008].

*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 [Stenström, Andersen, 1996]. 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 [Meyer, 2002], тогда как метаданные позволяют описывать использование языка в разных аспектах его функционирования.

### **Вопросы и задания для самоконтроля**

1. Дайте определение следующим понятиям: *конкорданс, корпусный менеджер, информационный запрос, интроспекция*.
2. Чем отличается понятие «корпусный менеджер» в широком смысле от понятия «корпусный менеджер» в узком смысле?
3. Что такое язык регулярных выражений?
4. Какие функции должен выполнять корпусный менеджер?
5. Почему нелингвистические корпусы (базы данных поисковых систем) можно рассматривать как корпусы текстов?
6. Каковы недостатки их использования в качестве корпусных менеджеров?
7. Что представляет собой внешняя и внутренняя статистика в корпусных службах?
8. Каковы недостатки корпусных менеджеров четвертого поколения?
9. Какие операции можно выполнить над конкордансом в системе Sketch Engine?
10. Как могут использовать корпусы разные специалисты?

## Часть 4

# ЛИНГВИСТИЧЕСКИЕ ИССЛЕДОВАНИЯ НА БАЗЕ КОРПУСОВ

## Глава 10. 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 связаны с необходимостью 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 [Герд, 2006].

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, как они применяются в разных контекстах и какова их сочетаемость. В частности, языковые исследования, основанные на эмпирических данных, проводятся лексикографами, работающими над оксфордскими словарями ан-

глийского 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. Реппена «Corpus Linguistics. Investigating language structure and use» [Biber, Conrad, Reppen, 1998] (далее — Corpus Linguistics) выделяется шесть основных вопросов, стоящих перед исследователями-лексикографами, действующими на основе корпусного подхода:

- Какие значения ассоциируются с конкретным словом?
- Какова частотность слова относительно других близких к нему слов?
- Какие нелингвистические модели имеет данное слово (по отношению к регистрам, историческим периодам, диалектам)?
- Какие слова обычно встречаются вместе с данным словом и каково распределение этих сочетаемостных последовательностей в разных регистрах?
- Как распределены смыслы и типы использования слова?
- Как используются и по-разному распределяются слова, кажущиеся синонимичными? [Biber, Conrad, Reppen, 1998]

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, ассоциирующиеся со словом. Корпус также предоставляет лексикографам разнообразную статистику.

### 10.1. 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 сотни тривиальных [Баранов, 2007].

Например, для слова *deal* из восьмимиллионного подкорпуса — корпуса **Лонгман-Ланкастер** выдается более 1500 употреблений, что усложняет задачу сгруппировать разные смыслы слова или распортировать их по важности. В таком случае приходится использовать дополнительные инструменты. Так, большинство программ конкордансеров могут создавать частотный список слов, который обычно представляется по убыванию частоты или в алфавитном порядке.

Приведем из учебника *Corpus Linguistics* пример выявления значений, ассоциируемых со словом *deal* в английском языке [Biber, Conrad, Reppen, 1998].

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 слово *look*, которое встречается около 500 раз на 1 млн слов. Более правильное решение в таком случае — 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. 10.1 показан частотный список, выданный программой TACT (Text-Analysis Computing Tools). Он показывает распределение

ние грамматических форм слова *deal* в аннотированном корпусе Ланкастер-Осло-Берген. Грамматическая категория каждого слова следует непосредственно за словом после символа «подчёрк». Так, слово *deal* встречается как существительное в единственном числе (граммема *nn*) 115 раз, как имя собственное (*np*) — 1 раз, как глагол (*vb*) — 66 раз. 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 10.1. Частота форм слова *deal* в аннотированном корпусе Ланкастер-Осло-Берген

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

### 10.1.1. 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 существительное 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. 10.2 включает абсолютные (сырые) частоты (*raw counts*) и нормированные частоты (*normed counts*) в пересчете на 100 тыс. словоупотреблений.

Таблица 10.2. Частотность существительного *deal* в определенных регистрах, нормированная на 100 тыс. слов

Регистр \\ Частота	Примерное количество слов в подкорпусе	Абсолютная частота для <i>deal</i>	Нормированная частота для <i>deal</i> (на 100 тыс. слов)
Репортажи прессы	88 000	14	15,9
Обзоры прессы	34 000	4	11,8
Передовицы	54 000	4	7,4
Религиозная литература	34 000	5	14,7
Научная литература	160 000	16	10,0
Научно-популярная литература	88 000	11	12,5
Беллетристика	154 000	24	15,6
Художественная проза	58 000	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. 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 раз, 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. Следовательно, только нормированные подсчеты обеспечивают достоверные основания для сравнения по регистрам.

Когда случаи встречаемости существительного *deal* распределены по регистрам, проблема размера корпуса для лексикогра-

фической работы становится еще более очевидной. Табл. 10.2 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.

Табл. 10.3 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.

Таблица 10.3. Частотность существительного и глагола *deal* в подкорпусах из двух регистров корпуса Лонгман-Ланкастер

Регистр	Частота Примерное количество слов в подкорпусе	Нормированная частота (на 1 млн слов)	
		Существительное	Глагол
Всего	4 000 000	90	119
Художественная проза	2 000 000	107	63
Научная литература	2 000 000	74	176

Во-первых, нормированные подсчеты показывают, что *deal/deals* как глагол лишь ненамного чаще встречается, чем *deal/deals* как существительное (119 слов на 1 млн словоупотреблений в сравнении с 90 словами на 1 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 1 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 1 млн слов).

Эти модели употребления *deal* высвечивают и другой важный момент в создании корпуса: корпус, ограниченный одним из регистров, не будет представлять язык во всей полноте. 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, 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.

#### 10.1.2. 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 ассоциирования с одним смыслом или значением. 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. 10.4 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 корпусных менеджеров.

Левые коллокаты — это слова, которые непосредственно предшествуют существительному *deal*. Например, из данных, представленных в табл. 10.4, 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 1 млн слов), затем следует прилагательное *good* (23 раза на 1 млн слов).

*Таблица 10.4.* Частотные коллокаты существительного *deal* в двух подкорпусах корпуса Лонгман-Ланкастер (5,7 млн слов)

Подкорпус	Нормированная частота (на 1 млн слов)
<i>Научная литература (подкорпус 2,7 млн слов)</i>	
<i>Левые коллокаты</i>	
great	45
good	23
<i>Правые коллокаты</i>	
of	39
more	7
in	3
to	3
<i>Художественная проза (подкорпус 3 млн слов)</i>	
<i>Левые коллокаты</i>	
great	40
good	28
the	8
big	3
<i>Правые коллокаты</i>	
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. 10.4, которая показывает только коллокаты, встретившиеся хотя бы 3 раза на 1 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 1 млн слов). Частотность следующего коллоката намного меньше, ср.: *more* (7 раз на 1 млн слов), *in* и *to* (3 раза на 1 млн слов). Итак, существительное *deal* чаще всего имеет значение количества, как в словосочетаниях *a good/great deal of*.

Анализ смыслов слова можно проверить, просмотрев конкордансы, 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. 11 показывает, что словосочетания в художественной прозе имеют интересные сходства и различия со словосочетаниями в научной литературе. 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 1 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 1 млн слов и *big* встречается 3 раза на 1 млн слов. Остальные 37 коллокатов встречаются не чаще двух раз на 1 млн слов.

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 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 четыре. 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. 10.5 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. Во-первых, 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 до 23-й дефиниции в двух словарях. Во-вторых, анализ коллокатов обнаружил относительно частотное значение, не отмеченное в этих словарях, — использование *big deal* в значении «незначительность». Наконец, во всех пяти словарях регистрационные отличия не принимаются во внимание, хотя более поздние словари, созданные на основе корпусных данных, начинают учитывать важные регистрационные модели.

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, то

Таблица 10.5. Словарные дефиниции существительного *deal*

Словари Значение	Webster's Encyclo. 1989	Webster's Third 1981	Chambers 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/arrangement	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

есть то, что корпус текстов не является самодостаточным [Рыков, 2002]. 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 нечасто (кроме карточных игр!). Этот пробел высвечивает важность больших представительных корпусов для лексикографической работы. Он также показывает, что 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 анализ, 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, принимая во внимание их регистровую отнесенность.

### 10.1.3. Слово *deal* как глагол

Значения *deal* как глагола в учебнике *Corpus Linguistics* рассматриваются на примере словосочетаний 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». Просмотр всего списка конкордансов показывает, что это значение намного чаще встречается в научной литературе, и в этом регистре коллокация *deal with* имеет большую 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 deal with.*

Наконец, коллокация *deal with* имеет значение взаимодействия 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 разными семантическими категориями (предмет разговора, проблемы). 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.

## 10.2. 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.

### 10.2.1. Распределение по регистрам синонимичных английских прилагательных *big*, *large* и *great*

Авторы учебника *Corpus Linguistics* рассматривают синонимичные английские слова *big*, *large* и *great*. Тезаурусы 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. 10.6 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, полученные абсолютные частоты могут не отражать 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 10.6. Частотное распределение *big*, *large*, *great* в подкорпусе объемом 5,7 млн с/у корпуса Лонгман-Ланкастер

Слова \ Частота	Ненормированные частоты	Нормированные на 1 млн с/у частоты
<i>Весь подкорпус (5,7 млн с/у)</i>		
big	1319	230
large	2342	408
great	2254	393
<i>Научная литература (2,7 млн с/у)</i>		
big	84	31
large	1641	605
great	772	284
<i>Художественная проза (3 млн с/у)</i>		
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 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 (490 и 408 на 1 млн словоупотреблений), тогда как *large* имеет намного более низкую частоту (232 раза на 1 млн словоупотреблений).

Из табл. 10.6 видно, что существует огромная разница в употреблении этих слов в двух регистрах. С одной стороны, *big* встречается более чем в 10 раз чаще 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. 10.7 и 10.8 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 в каждом регистре оказываются только 10 первых в списке коллокатов, которые встречаются чаще одного раза на миллион словоупотреблений (отсюда незаполненные столбцы для *big* и *large*).

Таблица 10.7. Десять 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

В обоих регистрах коллокаты *big* показывают, что это прилагательное чаще всего используется в отношении физического размера, однако 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.*

*Таблица 10.8. 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)*

<i>big</i>		<i>large</i>		<i>great</i>	
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

Вторая частотная пара в научной литературе — *big traders* — встретилась только 3 раза в одном тексте об экономическом развитии в западной Африке, где 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 отно-

шение 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 for 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. п. 13.3).

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 встречаются в художественной прозе, как *big*, и поэтому эти коллокаты реже употребляются с *large*, чем с *big*. Например, *big man* встречается 9,6 раза на 1 миллион, *large man* встречается всего 2 раза на 1 миллион. Подобно этому, *big house* встречается 7,6 раза на 1 миллион, а *large house* — всего 3 раза на 1 миллион словоупотреблений.

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, что проявляется в коллокации *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 по смыслу 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 коллокатами *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 коллокации с *great* в художественной прозе показывают, что у него гораздо больший спектр смыслов. Например, *great man* встречается 6,6 раз на 1 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 примерах: *a great, black bird; his great black moustache.*

Сочетаемостная пара *great burrow* (нора) также связана с физическим размером. Эта пара 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 усиливатель (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 Cad-dies parked out by the hydrangea bushes.* *Great* также употребляется в значении усиления 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 сочетания *great aunt* есть специализированное значение фамильного родства: *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 эта краткая информация показывает, что корпусный анализ позволяет выделять определенные модели, существенные для использования этих синонимичных слов. *Big* чаще всего относится к физическому размеру, *large* — к количеству, *great* также относится к количеству, особенно в сочетании *great deal*, но у этого прилагательного более широкий спектр значений, от усиления до терминов родства.

Три предпочтительных 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 у прилагательного *great* гораздо больше разных значений по сравнению с научной литературой.

Таким образом, синонимичные прилагательные совершенно не эквивалентны по своим значениям и по сочетаемости, когда реальные модели их использования анализируются на большом эмпирическом материале. Основанный на корпусных данных анализ можно использовать для показа того, что каждое прилагательное имеет собственные предпочтительные коллокаты, различные предпочтительные значения и различное распределение по регистрам.

### 10.2.2. Удаленные коллокаты *large*

Словам не обязательно примыкать друг к другу, чтобы их можно было ассоциировать друг с другом. Два слова могут иметь тенденцию совместной встречаемости, даже если между ними расположено несколько слов. Чтобы проиллюстрировать полезность рассмотрения разрывных сочетаний, здесь будут описаны сочетаемостные модели с коллокатом на втором месте справа от *large*, то есть синтагмы типа *large <существительное> <коллокат>*.

Табл. 10.9 показывает частотные коллокаты, встречающиеся в двух словах вправо от *large*. Коллокация *large N of* является частотной как в научной литературе, так и в художественной прозе. Действительно, эта коллокация так часто встречается, что ее уже можно рассматривать как рамку (*frame*), допускающую подстановку целого спектра слов, таких как *large amount of*, *large proportion of*, *large group of*.

Таблица 10.9. Наиболее частотные коллокаты слова *large* в позиции второго слова 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

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

- 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*;
- маркировка части сущности, большей по размеру: *a large proportion of, a large part of, a large sample of, a large fraction of, a large segment of*.

В этом случае самая сильная сочетаемостная ассоциация для *large* встречается со служебным словом *of* на расстоянии «плюс один». Однако исследование существительных в составе этой коллокации показывает, что эта рамка типично используется для маркировки количества или части большего целого.

Табл. 10.9 идентифицирует много интересных сочетаемостных ассоциаций. Например, модель *large X eyes*. Ассоциация особенно интересна, потому что она встречается более часто, чем сочетаемостная пара *large eyes; large X eyes* встречается 4,3 раза на 1 миллион слов, тогда как *large eyes* — только 1,6 раза на 1 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*.

## Глава 11. 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* [Biber, Conrad, Reppen, 1998] пути решения этой задачи проиллюстрированы на примере распределения 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 и в размеченных корпусах, и в неразмеченных. Многие корпусные менеджеры используют механизм усечения (*wild cards*) и позволяют пользователю искать определенные префиксы и суффиксы, например *-in-*, *-ment-*.

### 11.1. Распределение и функции номинализаций

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

#### 11.1.1. Анализ распределения номинализаций по регистрам

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, производилась также ручная обработка, чтобы исключить 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 подкорпусами корпуса Лонгман-Ланкастер. Третий регистр — устная речь — представлен корпусом Лондон-Лунд (объемом 500 тыс. словоупотреблений). Все частоты здесь нормированы на 1 млн слов текста.

Табл. 11.1 показывает частотные распределения для номинализаций по трем регистрам. В текстах художественной прозы и устной речи отмечаются близкие частотности, а в текстах научной лите-

туры 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 обрабатывает каждую индивидуальную номинализацию и выдает конкордансы 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.

Таблица 11.1. 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

Специфические номинализации, часто встречающиеся в регистре, зависят от тем, затронутых в текстах корпуса. Так, в научной литературе встречаются шесть номинализаций существенно чаще других, с частотами более 500 на 1 млн слов: *movement* (почти 900 случаев встречаемости на 1 млн слов), *activity*, *information*, *development*, *relation* и *equation*. Напротив, ни одна из этих 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* — 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. Ху-

дожественная проза и устная речь 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, поэтому в них 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 и значением номинализаций. Научная литература намного чаще говорит о статической номинализации, в то время как художественная проза и устная речь описывают подобные действия конкретных людей с помощью глаголов и прилагательных.

### 11.1.2. 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. 11.2, не показывая сами частоты, демонстрирует относительные пропорции номинализаций с каждым суффиксом в каждом регистре.

Мы видим, что:

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

- Суффикс *-ment* встречается в большем количестве в устной речи и художественной прозе, чем в научной литературе.
- Суффикс *-ness* чаще встречается в художественной прозе, чем в каждом из двух других регистров.

Таблица 11.2. 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 %

Хотя нет абсолютных правил, отвечающих за выбор суффиксов номинализаций, тщательный анализ каждого типа выявляет определенные системные различия в значении, проливая свет на модели распределения, показанные в табл. 18. 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 менталь-

ные 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 обычно 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 registros, отражая первостепенные коммуникативные функции регистров.

## 11.2. 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, характеризуя стили определенных авторов, исследователи иногда использовали

сравнительные подсчеты существительных и глаголов. Такие подсчеты, по мнению авторов учебника *Corpus Linguistics*, могут также показать разницу по регистрам.

### 11.2.1. Частотность грамматических категорий

В морфологически размеченном корпусе подсчет частоты существительных и глаголов — относительно легкое дело. Более сложным, если говорить об английском языке, является вопрос о том, что именно относить к глаголам и существительным (в первую очередь это проблема разметки):

1. Считать ли существительными слова, которые употребляются в роли определения последующего существительного? Например, должны ли словосочетания типа *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, словосочетание *grasshopper ecology* содержит два отдельных референта — кузнечики (*grasshoppers*) и экология (*ecology*). В этом смысле они отличаются от фраз с прилагательным в качестве определения. В описываемом исследовании существительные, которые определяют другие существительные, рассматриваются как 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*. Так как нужно подсчитать слова, непосредственно относящиеся к предметам, местоимения при этих подсчетах не учитываются.

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 ни-

какого лексического содержания. Вместо этого они служат только для маркировки аспектуального значения (*perfect* или *progressive*) либо 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 и для исследователя. Для пользователя важно 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 (*noiety*) является регистр. Табл. 11.3 показывает соотношение «существительное/глагол» при использовании трех разных способов подсчета существительных и глаголов.

Таблица 11.3. Соотношение «существительное/глагол» в трех регистрах

Регистр Категория		Научная литература	Художественная проза	Устная речь
A	Все существительные и глаголы	2,2 : 1	1,2 : 1	1,2 : 1
B	Все существительные и глаголы, за исключением вспомогательных	2,9 : 1	1,5 : 1	1,6 : 1
B	Существительные, за исключением использованных в роли определения, и глаголы, за исключением вспомогательных	2,5 : 1	1,3 : 1	1,3 : 1

Строка А учитывает все существительные и глаголы, строка Б не учитывает 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. Обратим внимание, что количественные соотношения одинаковы 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 полученные по всем регистрам модели одинаковы, хотя точные соотношения в регистрах варьируют.

#### *11.2.2. Сравнение соотношения «существительное/глагол» по регистрам*

Авторами учебника *Corpus Linguistics* самым подходящим способом подсчета существительных и глаголов признан подход Б (см. табл. 11.3), то есть исключение местоимений и вспомогательных глаголов из подсчетов. Соотношение «существительное/глагол» в научной литературе намного превысило соотношение в других регистрах (2,9 против 1,5 и 1,6), что может быть интерпретировано также, как и результат исследования номинализаций (см. п. 11.1). 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 соотношение «существительное/глагол». Так, в примере неформальной беседы больше местоимений, чем существительных. 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 и описания 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 для чисто научных. Интуиция носителя языка не всегда является надежной в предсказании того, какой вариант предпочтительнее другого. Только основанное на корпусных данных исследование реальных текстов подходит для выявления этих моделей [Biber, Conrad, Reppen, 1998].

## Глава 12. 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 экспрессивистическими — 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) [ЛЭС]. Элементы дискурса: излагаемые события, их участ-

ники, перформативная информация и «не-события», то есть обстоятельства, сопровождающие события; фон, поясняющий события; оценка участников событий; информация, соотносящая дискурс с событиями. 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 [Virtanen, 2008], две последние, так как именно они выигрывают от широкомасштабных исследований, возможных с помощью корпусов, поскольку вариативность предполагает устоявшиеся и новые правила, а коллокации предоставляют доступ к осозаемым 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:

- 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;
- 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.

## 12.1. Характеристики референциальных выражений

### 12.1.1. Распределение референциальных выражений по регистрам

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) с целью найти ответ на следующие вопросы:

- Какие факторы влияют на выбор между существительными и местоимениями в тексте?
- Какие существительные представляют «данную» (или «известную») информацию, а какие представляют «новую» информацию?
- 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 artikel *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 или позже, либо выходить за рамки текста, но входить в контекст дискурса.

Выделяют три 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, то есть для ее правиль-

ного 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 «отсылаочных выражений». В учебнике *Corpus Linguistics* рассматриваются два примера из новостного репортажа и неформальной беседы, в которых именные конструкции выделены курсивом:

**Пример 4.** 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 5.** 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 касается формы именных конструкций. В примере из новостного репортажа 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 эк-

зофорическую 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.

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

- статус информации: известная, новая;
- тип референции для известной информации: анафорическая, экзофорическая или выводимая;
- форма выражения для анафорической референции: местоимение, синоним или повтор;
- 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 4) 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 информацию, вводя сущность, которая уже была упомянута. Так, в первом предложении местоимение *it* употреблено дважды, чтобы обозначить известный референт — компанию *Thortec International Inc.*

Выражения, вводящие известную информацию, представляют три типа референциальных отношений. Многие из таких выражений являются анафорическими, то есть относятся к человеку или объекту, уже упомянутому в тексте, — антецеденту. Так, антецедентом для местоимения *it* в первом предложении является *Thortec Inter-*

*national 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 5) местоимения *I* и *you* прямо указывают на говорящего и адресата. *The back door, the car and the paper shop* относятся к физическим объектам, присутствующим в расширенной физической ситуации, которая понятна обоим участникам разговора. Такие референты называются экофорическими. 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 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 consulting 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 (а это очень мало) может быть весьма

затратным по времени, поэтому для изучения характеристик референциальных выражений целесообразно использовать корпусный подход.

### **12.1.2. 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:

- регистр, который предварительно указывается в начале каждого текста и не вовлекается в последующий анализ;
- форма именной конструкции (местоимение или существительное), определяющаяся на основе процедуры аннотирования;
- статус информации (новая или известная), причем местоимения автоматически рассматриваются как известная информация, а для каждого существительного производится проверка, выявляющая, 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;
- 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;
- тип выражения (синоним или повтор существительного), если представлен анафорический тип референции, выраженный существительным;

- расстояние между антецедентом и референциальным выражением, 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. 12.1 приводится пример работы программы, показывающий, как коды могут быть приняты или отредактированы.

```
*** Code Check *** (processing file 00057 . TEC; word 366)
impressive that quantum mechanics can take that in its
stride. The problems of interpretation cluster around
two issues; the nature of reality and the nature of
measurement. Philosophers of science have latterly
been busy explaining that science is about correlating
phenomena or acquiring the power to manipulate
= = => them.

They stress the theory - laden character of our pictures
of the world and the extent to which scientists are said
to be influenced in their thinking by the social factor of
the spirit of the age . Such accounts cast doubt on whether
an understanding of reality
Automatically assigned code is: REF= ANAPHORIC
ALTERNATE CODES ARE:
1) REF= ANAPHORIC          2) REF= EXOPHORIC
3) REF= INFERRABLE         4)
5)                         6)
7)                         8)

Type number 1-8 to select alternate code
Push <ENTER> to accept code; * to terminate file;
c for more context
```

Рис. 12.1. Пример работы интерактивной программы кодирования референциальных выражений

Референциальное выражение (*them*), которое подлежит кодированию, представлено в контексте и обозначено стрелкой. Под текстом примеров приведены автоматически присвоенный код (*anaphoric*) и альтернативные варианты кодов. Когда все именные конструкции проанализированы, коды записываются в тексте так:

```
<<<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 важно использование корпусного подхода для выявления характеристик дискурса, присущих тому или иному регистру [Biber, Conrad, Reppen, 1998; Толпегин, 2008].

## 12.2. 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 [Leech, 1999]. Объектом проведенного на корпусе исследования была грамматика обращений, понимаемых как субстантивные свободно присоединяемые элементы, не являющиеся членами предложения и 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. 12.1), а также различные отношения между типом функции (привлечение внимания, указание на адресата, усиление социального взаимодействия) и позицией обращения в предложении.

Таблица 12.1. Место обращения в предложении

Место в предложении	Кол-во, %	Пример
Конец предложения	68,00	Come on, Sam.
Начало предложения	11,50	Doug, do you want some more ice-cream?
Отдельное расположение	11,25	Mom!
Внутри предложения	9, 25	What have we lost at home, Paulie, this season?

Автор полагает, что ему удалось получить новые результаты:

- Люди используют обращение *sir* в разговоре с официантами, что, возможно, объясняется тенденцией в направлении демократизации общества.
- Обращения чаще всего встречаются в конце предложения, что может быть объяснено большей важностью их социальной функции по сравнению с другими, включая функцию привлечения внимания.

### 12.3. Пример исследования дискурса на материале речевого корпуса

На материале русского языка проводилось исследование энантиосемии — совмещения в слове противоположных значений, «внутренней антонимии» [Маркасова, 2008]. Рассмотрение этого явления на материале звукового корпуса русского языка повседневного общения *Один речевой день* (ОРД) (см. п. 7.2.2.5) привело автора к мысли о том, что есть прежде не выявлявшийся лингвистами тип энантиосемии, часто используемый в разговорной речи, — 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 (*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-, -очек-*:

- Але-о! / Привет / что все?/ закончилось / у вас закончилось /что случилось?/ почему? / ты где? то есть она сейчас уже ушла? а спроси у Иры / да-а / я же тебе сказала / подойти к Ирине и спросить / попросить помочи / 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 помочи найти Нину Филипповну / и отзвонись мне / пока.
- 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.
- 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? Зачем? <...> проверка по словам. <...> Какие слова? Что / и сейчас забыла? И сейчас забыла / какие слова? <...> Из какой лексики / **Лизонька**?
- **Лизонька** / это Марина Викторовна ваша / такое дурацкое слово употребляет «лексика» / оно дурацкое / **Лизочка** / Лексика — это всего-навсего слова. <...> Лексику вы учите / бред какой.
- **Моя девочка.** <...> Ну что делать-то с этой двойкой / **моя хорошая**.

Была выдвинута гипотеза о том, что особенности интонаирования 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 (группе из 30 студентов и школьников) было предложено прочитать расшифровку и ответить на 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 прокурорским 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». При всем разнообразии оценок поведения Лизы и ее матери реплики отражают одно: в прослушанных фрагментах участники эксперимента ощутили нечто неприятное, раздражающее, что 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 с (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 об-

разом, 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 [Крейдлин, 2000]. В отличие от нормального (неагрессивного) 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 снижается эмоциональное напряжение говорящего за счет близких людей [Маркасова, 2008].

Данный пример исследования показывает, что на материале речевого 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.

## Глава 13. Корпусные методы исследования

Корпусными называют методы, которые используются при работе с корпусами текстов. К ним относятся разнообразные методы сбора, аннотирования и осуществления поиска в корпусе, а также обработки корпусных данных [Lüdeling, Kytö, 2008]. Из данного определения вытекает несколько обособленных групп методов корпусной лингвистики:

- методы, применяющиеся при создании корпусов текстов;
- методы, применяющиеся в процессе поиска в корпусе (автоматизированное извлечение информации);
- методы обработки полученных данных.

Далее обозначим некоторые из этих методов в соответствии с выделенными группами и рассмотрим, как они применяются в современных исследованиях русского языка.

### 13.1. Применение корпусных методов сбора, обработки и аннотирования текстового материала

Методы сбора и аннотирования текстового материала являются собственно методами корпусной лингвистики, разрабатываемыми со временем создания первых корпусов текстов, то есть с 1960-х годов. В основе аннотирования (разметки), в частности по частям речи, лежали масштабные исследования в области автоматической обработки естественного языка (NLP), конечной целью которых было создание искусственного интеллекта.

Далее следуют некоторые примеры корпусов русского языка, находящиеся на разных стадиях разработки, сбора текстового материала и аннотирования.

#### 13.1.1. Корпусы делового языка

Фактическим идеологом корпусной лингвистики в СССР был академик А. П. Ершов (1931–1988), один из первопроходцев отечественного программирования. Корпусная русистика зародилась тогда, когда А. П. Ершов предложил создать корпус русской деловой прозы [Ершов, 1979]. Однако до последнего времени корпус, адекватно и системно представляющий официально-деловой функциональный стиль русского языка, по-прежнему не создан. В связи с этим поставлен вопрос о создании **Официально-делового корпуса русского языка (ОДКРЯ)**, включающего не образцы жанров официального дискурса, а конкретные юридические документы (послепетровского периода — XVIII–XXI вв.). Потребность в таком корпусе продиктована несистемной представленностью юридических и законодательных текстов в существующих корпусах русского языка: они не позволяют проследить состав юридической терминологической системы и особенности формирования и изменения официально-делового русского языка. Важной представляется возможность вести поиск как по словам и конструкциям, так и по названиям документов. ОДКРЯ, включающий законодательные документы, позволит наблюдать за актуализацией единицы в специальных и неспециальных текстах и сохранять связь диахронного и синхронного подходов [Крылов, Фролова, 2017].

Если этот корпус пока находится в стадии разработки, работа над другими корпусами русского языка ведется уже несколько лет. Кор-

пус текстов российских правовых актов RusLawOD (<https://github.com/irlcode/RusLawOD>) сформирован Институтом проблем право-применения при Европейском университете в Санкт-Петербурге на основе документов, доступных на Официальном интернет-портале правовой информации ([www.pravo.gov.ru](http://www.pravo.gov.ru)) [Савельев, 2018].

Корпус состоит из документов в формате XML, содержащих тексты и метаданные документов. Часть документов на портале прошла официальное электронное опубликование, и их тексты получены после оптического распознавания сканов графических страниц (OCR). Остальная часть получена из HTML-текстов раздела «Законодательство России» портала [pravo.gov.ru](http://pravo.gov.ru). В корпус текстов включены только первоначальные редакции правовых актов на момент принятия, а также последующие акты об изменениях. Консолидированные версии документов с внесенными изменениями не представлены. За 1991–2017 гг. объем корпуса текстов составлял 458 884 документа. Всего в корпусе более 600 млн токенов.

С использованием данного корпуса методами компьютерной лингвистики проведено исследование, в котором была проанализирована динамика изменения лексического и синтаксического качества текстов правовых актов [Кучаков, Савельев, 2018]. На основе исследования сделаны выводы о том, что в России наблюдается ухудшение качества текстов федеральных и региональных правовых актов для восприятия — уменьшение лексического разнообразия, усложнение структуры предложений. В последние годы эта тенденция усилилась. Отмечено, что наиболее сложные конструкции предложений встречаются в текстах Конституционного суда РФ, а также органов власти, связанных с финансово-бюджетной сферой регулирования.

### **13.1.2. Корпусы диалектов**

Отдельные говоры представляют важнейшие диалектные типы русской речи. При этом в соответствии с принципом пропорциональности текстовая база корпуса каждого отдельного говора должна быть направлена к моделированию коммуникации в данном говоре, отражая важнейшие типы и формы диалектной речи, социальную дифференциацию носителей говора, жанрово-тематическую структуру диалектного общения [Крючкова, Гольдин, 2017]. Эти принципы лежат в основе создания **Саратовского диалектологи-**

**ческого корпуса** (СарДК) (с 2008 г.), в котором разрабатываются подкорпусы отдельных говоров разных типов (в настоящее время это подкорпус северорусского говора с. Мегра Вытегорского района Вологодской области, подкорпус среднерусского окающего говора с. Белогорного Вольского района Саратовской области, подкорпус среднерусского акающего говора с. Земляные Хутора Аткарского района Саратовской области). Специфика диалектного материала требует контекстов большей протяженности, чем в корпусе стандартных текстов, и возможности получения целого текста. Именно этими положениями определяются параметры выдачи по запросу в СарДК, где минимальной выдачей является абзац, то есть структурно-семантическое целое, а максимальной — целый текст как политематическое и полижанровое единство обычно значительной протяженности. Возможность выдачи текстовых фрагментов по тематическому или жанровому критериям (создание пользователем тематических или жанровых подкорпусов) обеспечивается реализуемыми в СарДК тематической и жанровой разметками каждого выделенного на формальной основе диалектного текста. Говоря о метаразметке текстов в СарДК, следует упомянуть сведения о конкретной ситуации записи текста (в доме информанта, в поле, в огороде, в лесу и т. п.), об адресатах речи, об упоминаемых в тексте лицах, о времени описываемых в тексте событий, а также о важности достаточно подробных сведений об информанте (местный, приезжий, откуда приехал, как давно живет в данном населенном пункте, род занятий, конфессия и др.). В перспективе речь идет о дополнении метаразметки фонетическим комментарием к тексту [Крючкова, Гольдин, 2015; 2017].

### **13.1.3. Корпус устной речи «Один речевой день»**

Еще один корпус устной речи, «Один речевой день» (ОРД), представляющий речь жителей большого российского города, продолжает создаваться в СПбГУ. Программа аннотирования в системе ELAN дает возможность реализации многоуровневой лингвистической разметки. На данном этапе разработки корпуса, позволяющего фиксировать и осуществлять мониторинг естественной русской речи, ставится задача разработки расширенного шаблона многоуровневого аннотирования речевого материала, включающего в себя дополнительные уровни, характеризующие тематику разговоров

и эмоциональную окраску речи. Запись фрагментов повседневного бытового общения на русском языке, выполненная в естественных условиях, осуществлялась в Санкт-Петербурге в 2007–2016 гг., ее общий объем составляет 1250 часов звучания (2800 коммуникативных макроэпизодов). По состоянию на июнь 2017 г. получены текстовые расшифровки 17% звукозаписей корпуса (480 макроэпизодов), в них насчитано 1 млн словоупотреблений [Шерстинова, 2017]. Важным требованием к тематическому аннотированию при работе с мультимедийным контентом является сегментация аудиофайлов на фрагменты, относительно однородные по теме разговора (микроэпизоды). При этом введение расширенного шаблона аннотирования может привести к необходимости пересмотра (коррекции) уровня микроэпизодов для уже расшифрованного подкорпуса звукозаписей, чтобы привести речевой материал к единому формату представления данных. Для звукозаписей корпуса с текстовыми расшифровками, уже отсегментированных на тематически однородные микроэпизоды, можно провести тематическое аннотирование с использованием статистических методов автоматического извлечения ключевых слов. Однако, учитывая высокую эллиптичность повседневной бытовой речи, когда «тематические слова» в разговоре довольно часто опускаются, эффективность применения таких методов требует специальной проверки. Кроме того, абсолютное большинство существующих в настоящее время программ ориентировано на письменные тексты, состоящие из предложений, поэтому «квантами» автоматического анализа для извлечения ключевых слов являются предложения. В спонтанной устной речи членение на единицы, соответствующие предложениям, далеко не всегда можно выделить однозначно, поэтому при обработке расшифровок устной речи имеет смысл ориентироваться не на предложения, а на другие показатели — длительные (разграничительные) паузы в потоке речи или определенное количество словоупотреблений.

При тематическом аннотировании речевого материала некоторую сложность может вызвать одновременное обсуждение группой собеседников нескольких тематически далеких тем, а также аннотирование тех ситуаций, когда анализируемый коммуникативный эпизод состоит из нескольких параллельных разговоров (например, застольные разговоры или разговоры в офисе). Тематическое аннотирование корпуса ОРД даст возможность проводить поиск данных по теме бытового общения (разговор о здоровье, о личных

взаимоотношениях, о работе, о новых гаджетах и др.). Помимо оптимизации поисковых возможностей тематическое аннотирование аудиозаписей имеет и другую важную цель, которая связана с исследованием тематического разнообразия повседневных бытовых разговоров, выяснением того, какие темы чаще других становятся предметом устного общения у жителей большого российского города в начале ХХI в. [Шерстинова, 2017].

Для корпуса ОРД разработаны также принципы многоуровневого аннотирования микроэпизодов, подразумевающие заполнение следующих уровней аннотации: 1) доминантной прагматической коммуникативной задачи; 2) эмоционального фона; 3) обобщенного типа коммуникативного сценария (бытового разговора, профессионального разговора, клиент-сервис коммуникации, обучения и др.); 4) социальной роли информанта (отца, сына, коллеги, друга); 5) места коммуникации (дома, офиса, кафе); 6) формальной оценки успешности коммуникации; 7) количества векторов коммуникации. По результатам пилотного прагматического аннотирования корпуса ОРД получена статистическая информация о функциональной активности разных типов речевых актов для отдельных говорящих, для отдельных коммуникативных сценариев и в целом (на материале шести коммуникативных макроэпизодов в объеме 2250 речевых актов). Как и следовало ожидать, чаще всего в повседневном общении используются презентативы, которые составляют 38,53% всей речевой коммуникации, на втором месте по частоте идут разнообразные регулятивные формы — 12,36%, вердиктивы-валюативы составляют 11,15%, директивы — 6,67%, этикетные речевые акты — 4,13%, паралингвистические формы — 3,55%, экспрессивы-эмотивы — 3,42%, комиссивы — 2,58%, вердиктивы-суппозитивы — 2,53% [Шерстинова, 2015].

Вопросы разметки в применении к корпусу ОРД (1 млн с/у в расшифровках) поднимает О. В. Блинова при анализе семантических и прагматических свойств побудительных (в частности, императивных) реплик типа ну говори! ну озвучивай / милая! \*П быстрее! Автор основывается на наличии и характере речевых реакций, вызванных этими репликами. При разметке учитываются три типа информации: 1) информация о внутренних свойствах и диалогических функциях реплик (*resp. intra-utterance features*); 2) информация о диалогическом контексте (*resp. inter-utterance features*); 3) информация о сходстве между стимульными и реактивными репликами. Для

целей исследования оказалось целесообразным экспортировать тексты из программы ELAN в CLAN. В таком случае транскрипт оказывается разделенным на строки, а каждая строка соответствует одной реплике (фрагменту реплики) и имеет порядковый номер. Деление на строки и присвоение номера происходят автоматически в результате экспорта из ELAN (\*.eaf) (который используется при создании ОРД) в CLAN (\*.cha). Существенно, что внутри каждого файла \*.cha присутствуют метаданные об участниках диалога, степень подробности которых можно варьировать [Блинова 2017].

#### **13.1.4. Учебный прагматический корпус**

При разработке Учебного прагматического корпуса был применен другой подход к проблеме разметки, поскольку в классификации прагматических явлений и представляющих их теорий нет единства [Тимофеева, 2017]. Для многих прагматических составляющих текстовые маркеры отсутствуют. Следовательно, выбор отображаемых в корпусе явлений представляет собой нетривиальную задачу. В имеющихся прагматических корпусах русского языка фиксируются, как правило, типы речевых актов (как в ОРД) и коммуникативная организация. В учебном прагматическом корпусе предлагается сделать акцент на неоднозначности, в частности на таких типах явлений, как: 1) неопределенность области действия кванторного слова/отрицания/обозначения логической операции; неопределенность границ/ролей аргументов предиката; неопределенность антецедента анафоры, например: *Все тетради и книги в портфеле*: ‘(тетради и книги) в портфеле’ или ‘тетради и (книги в портфеле)’; *В Бердске штрафуют владельцев собак, гуляющих без намордников и поводков*; 2) неопределенность экзистенциального типа: существование/несуществование объекта/процесса/события, о котором говорится в тексте, например: *Он не обрадовал нас своим приходом*, где выбор значения зависит от того, что является пресуппозицией — пропозиция «Он приходил» (но это «нас» не обрадовало) или пропозиция «Он не приходил» (тем самым лишив «нас» случая порадоваться его приходу); 3) неопределенность приписывания тексту прагматических пресуппозиций, по-разному разрешаемая разными людьми, например: — *Послушай, Клэр, не надо все-таки бросаться луковицами в мистера Брэддока.* — Так я сходила наверх и подобрала ее, — утешила Клэр свою хозяйку, — она уже в супе. (П. Г. Вудхаус). Первой реплике собеседницы приписывают

разные пресуппозиции: «Бросаться луковицами в людей нехорошо» и «Разбрасываться луковицами неэкономно». Во всех случаях прагматические составляющие рассматриваются как пропозициональные по своей природе, то есть как дополняющие явно высказанные в тексте пропозиции определенным набором неявных пропозиций (пресуппозиций). Разметка отражает альтернативные роли, границы, связи, ставя им в соответствие неявные пропозиции. Таким образом, помимо традиционной расстановки кодов языковых явлений разметка включает также дополнительные пропозиции (пресуппозиции), для представления которых должна быть выработана определенная стандартная форма [Тимофеева, 2017].

Основное назначение обсуждаемого прагматического корпуса — это накопление структурированного текстового материала, создаваемого людьми, изучающими определенные разделы прагматики. Получаемый в результате учебный корпус после внесения вошедших в него необходимых исправлений предполагается использовать для изучения прагматических явлений. Данный корпус можно отнести к иллюстративным корпусам текстов, описанным в п. 4.1. Единицей корпуса будут короткие фрагменты текстов (от одного до нескольких предложений), которые содержат одно или более входящих в них прагматических явлений, относящихся к рассматриваемым типам. Таким образом, корпус создаст среду с повышенной концентрацией изучаемых прагматических явлений.

### **13.2. Применение корпусных методов извлечения информации из русскоязычных корпусов текстов**

Методы, применяющиеся в процессе поиска в готовом корпусе текстов, в большинстве случаев определяются встроенным в корпусный менеджер поисковыми инструментами: поиск конкретных словоформ, построение конкордансов, метод извлечения ключевой лексики, выбор лексических и синтаксических коллокаций (повторяющихся, воспроизводимых конструкций) и др.

#### ***13.2.1. Корпусы и переводная лексикография***

По словам Л. Н. Беляевой, идея создания автоматизированных систем извлечения терминов из корпусов текстов насчитывает уже более 20 лет и в той или иной степени реализована в различных проек-

так [Беляева, 2015]. Под извлечением терминов понимается формирование лексических ресурсов различного типа: терминологических словарей, учебных словарей, переводных словарей, тезаурусов, лексико-семантических полей, списков именованных сущностей и т. п. Для этого могут использоваться как общеязыковые корпусы, так и специальные [Захаров и др., 2019]. Однако даже самая изощренная система извлечения терминов не дает окончательного результата для включения выбранных лексических единиц в переводной словарь, а предоставляет лишь удобно организованный и оперативно получаемый ресурс для работы терминолога или лексикографа.

В переводной лексикографии широко используются не только параллельные, но и сопоставимые корпусы. Обращение не к переводам текстов, выровненным по предложениям в параллельных корпусах текстов с их оригиналами, а к сопоставимым корпусам текстов, при организации которых возможна экспертная оценка текстов на сопоставляемых языках, вполне естественно, хотя и ставит дополнительный вопрос об их выравнивании. В случае сопоставимых текстов возможно только выравнивание по терминам (по кандидатам в термины), опирающееся на выявление характерных для обоих массивов корпуса однословных терминологических единиц и их сопоставление в качестве кандидатов в переводные эквиваленты, а также поиск устойчивых словосочетаний с этими однословными терминами в качестве ядер [Беляева, 2015]. Дальнейший сопоставительный анализ требует привлечения знаний из переводных словарей, позволяющих верифицировать выбранные пары терминов. Извлечение многокомпонентных терминов может при этом основываться на результатах автоматического синтаксического анализа на уровне функциональных сегментов — именных групп.

Выявление кандидатов в термины из корпусов однословных терминов может опираться на семантические характеристики этих слов, извлекаемые из различных автоматизированных баз данных и словарей систем машинного перевода. При использовании в качестве справочного массива словарей предметно-ориентированных систем, словарные статьи которых содержат синтаксические и семантические характеристики выявленных слов, такое соотнесение можно автоматизировать.

При создании терминологических баз данных предлагается опираться на те огромные словарные ресурсы, которые накоплены в раз-

личных системах машинного перевода, поскольку при отборе лексики в словари систем машинного перевода принимается во внимание не только терминологический статус лексической единицы (единицы перевода) — слова или словосочетания (машинного оборота), но и ее распространенность в конкретном языке для специальных целей. Автоматические словари в своей исходной части не являются словарями нормативными, поскольку в качестве заглавия словарной статьи в них используются все встречающиеся варианты номинации объектов, а перевод соответствует рекомендуемому для языка перевода. Тем самым автоматические переводные словари выполняют функцию нормализации терминологии только относительно языка перевода: словарь фиксирует все варианты термина (слова или словосочетания) на входном языке, сопоставляя их с нормативным (стандартизированным) вариантом на языке перевода. Так, например, словосочетаниям *test administrator*, *test supervisor* соответствует перевод *администратор теста*, словосочетаниям *bank of items*, *item bank* и универбату *itembank* соответствует перевод *банк тестовых заданий*.

Возможность опоры на универбаты при создании англо-русских и русско-английских переводных словарей с использованием информации из автоматических словарей Л. Н. Беляева рассматривает на примере именных машинных оборотов с ядерными словами *method* и *метод* из автоматических словарей для предметной подобласти лингводидактика. Обе эти лексические единицы относятся к общеначальной лексике, могут быть выявлены в любом научном тексте и, соответственно, в ядерной позиции могут служить опорой для установления кандидатов в термины. В анализируемых автоматических словарях зафиксировано 34 словосочетания с ядерным словом *method*, большинство из которых (20) состоит из двух компонентов и создано по модели A+N (табл. 13.1). Если рассматривать варианты перевода самого ядра, то только в одном из 34 именных терминологических словосочетаний при переводе английского слова *method* использован переводной эквивалент *способ*.

При рассмотрении русско-английского варианта оказывается, что при переводе слова *метод* возникает более сложная ситуация. В автоматическом словаре зафиксировано 65 именных словосочетаний с ядерным словом *метод* (табл. 13.2), что почти вдвое больше количества словосочетаний со словом *method* в ядре. При этом лексической единице *метод* в переводных эквивалентах соответствуют английские лексические единицы *method* (32), *approach* (2), *teaching*

(5), *instruction* (1), *analysis* (1), *learning* (1), *model(ling)* (2), *fashion* (1), *technique* (3), *program(me)* (1). Для остальных 16 словосочетаний нет прямого соответствия в переводных эквивалентах, что свидетельствует о более широком значении лексемы *метод* в русском языке.

Таблица 13.1. Словосочетания с ядерным словом *method*. Фрагмент

Английское словосочетание из АС	Русский переводной эквивалент в АС
acceptable word method	метод оценки с учетом слов, соответствующих контексту
adult method	принятый у взрослых метод
classroom method	метод организации образовательного процесса
comparative method	сравнительно-исторический метод
scoring method	способ подведения итогов

Таблица 13.2. Словосочетания с ядерным словом *метод*. Фрагмент

Русское словосочетание из АС	Английский переводной эквивалент в АС
аудиовизуальный метод	audiovisual method
аудиолингвальный метод	audiolingual method, audiolingualism
классический метод	classical approach
коммуникативный метод обучения	communicative instruction, communicative teaching
комплексный метод	companion analysis
метод «общины»	community language learning
метод анализа в глубину	depth first fashion
метод ведения родительского дневника	parental diary technique
метод математического моделирования	mathematical model

Данный пример убедительно показывает, что опора на информацию из автоматических словарей систем машинного перевода может дать полезную информацию при сопоставлении флексивных и аналитических языков [Беляева, 2015].

### 13.2.2. Веб-корпусы: *pro et contra*

Проблема поиска в корпусе текстов непосредственно связана с его репрезентативностью. Если под репрезентативностью понимать объем, то опыт работы с корпусами говорит о том, что во многих случаях объем Национального корпуса русского языка оказывается недостаточным для получения полноценных достоверных данных. Этого мало, как правило, для словосочетаний и совсем мало для фразеологизмов [Захаров, 2015а]. Такой объем совершенно не подходит и для диахронических исследований. Так, если сочетание *громкие аплодисменты* в НКРЯ встретились по одному разу в 1885, 1906, 1908, 1910, 1925, 1939–1940, 1959, 1963, 1998–2000, 2003 гг. и два раза — в 2001 г., то трудно делать какие-либо умозаключения на основе такого ничтожного количества данных [Захаров, 2015а]. Коллокации и коллигации для среднечастотных и низкочастотных слов реально изучать только на миллиардных корпусах. Создание корпусов — процесс трудоемкий и не очень быстрый, и как только была осознана необходимость создания корпусов большого объема, стало ясно, что потребности лингвистов и возможности корпусной лингвистики сильно расходятся. Многие лингвисты за решением своих задач обращаются к вебу, задавая запросы на языке поисковых систем в Интернете (см. п. 6.1), поэтому и родилась технология создания корпусов на базе веба, которая, казалось бы, решает проблему репрезентативности. Однако при работе с этими корпусами возникают новые проблемы, которые ставят вопросы как перед разработчиками, так и перед пользователями, — качество текстов веб-документов и проблема сбалансированности создаваемых корпусов. Покажем это на нескольких примерах [Захаров, 2015б].

В качестве материала и инструмента исследования были использованы НКРЯ (283,4 млн слов), корпус русских текстов ruTenTen 2011 системы Sketch Engine (14,5 млрд токенов) (<https://the.sketchengine.co.uk/>), корпусы русских текстов из семейства псевдо-параллельных корпусов Araneum Университета им. А. Коменского в Братиславе (<http://ucts.uniba.sk/>).

Тексты, взятые из Интернета, с большой долей вероятности содержат разного рода недостатки, влияющие в том числе на качество лемматизации и морфологического анализа и на качество лингвистического анализа: опечатки, орфографические вариации, ошибки капитализации, слова из других (похожих) языков, обрывки слов

(часто из-за переноса), имена собственные, экспрессивную лексику, новые слова, которые не удается правильно лемматизировать, например: *слушаю-с, хрюкмены, щаскакам, приве-е-ет, вылысыты-дыстко, иоаннутый, та-а-ак, триногометрия, советобоязнь, нью-вэйв, Архнадзор* и т. п.

Частотный словарь в корпусе ruTenTen насчитывает более десяти миллионов слов, в корпусе Araneum Russicum Maius — порядка 5 млн слов. Там представлены и ошибочные слова, и нелемматизированные словоформы, и много другое. В табл. 13.3 приведены примеры ошибочных лемм из такого словаря, точнее, того, что получается на выходе морфологического анализатора TreeTagger с частотами из корпуса Araneum Russicum Maius.

Таблица 13.3. Примеры ошибочных лемм с частотами из корпуса Araneum Russicum Maius

Ошибочные леммы	Частота
Януковичу	240
Януковичем	142
я-то	129
ЯНАО	83
яже	80
Яннушка	47
явл-ся	29
які	28
ямальцев	27
Яврэ	21
Якщо	20
Яффо	19
якудза	11
языцы	11
языкъ	11
явля	11

Окончание табл. 13.3

Ошибочные леммы	Частота
яться	10
ямобура	6
янь	5
языков-миф	1
языкоблудием	1
языкоблудивой	1
языко-христиан	1
языкк Ncmsan	1

Встает вопрос, насколько такой «грязный» словарь сказывается на качестве результатов.

На основе корпуса Araneum Russicum Minus для 1000 наиболее частотных слов, для которых лемма не распознана, была собрана сравнительно небольшая статистика. Среди нелемматизированных слов — неологизмы, сленг, а также новые слова, образованные с помощью разных префиксOIDов. Их неполный список насчитывает около 400 единиц. Для каждого префиксOIDа из этого списка была подсчитана абсолютная частота и количество употреблений на 1 млн слов (*ipt*) по НКРЯ и по корпусу Araneum Russicum Maius. Была составлена сводная таблица, и каждому префиксOIDу присвоен ранг в обоих корпусах. Видна разница между списками из двух корпусов, причем в корпусе Araneum большую частоту имеют такие префиксOIDы, как *Интернет-*, *веб-*, *евро-* и т. д.

Особенность веб-корпусов заключается в том, что они получаются путем неуправляемого кроллинга Интернета, повлиять на их сбалансированность невозможно. Более того, судить о конечной сбалансированности в терминах жанров и регистров вообще сложно, так как в веб-документах не просто отсутствует жанровая метаразметка, но и само понятие жанра применительно к ним должно быть изменено.

Эксперименты по выявлению общей корреляции разных веб-корпусов с традиционным НКРЯ путем сравнения словарей (в качестве словаря НКРЯ выступал новый Частотный словарь современного русского языка [Ляшевская, Шаров, 2009]) по 25 суще-

ствительным из средне-верхней частотной зоны показали заметное расхождение между корпусами в лексике: по коэффициенту Спирмена средний коэффициент корреляции равнялся 0,58, средний коэффициент Пирсона составил 0,87, сравнение по мере *t-score* показало результат 1,57. За параметры лексических единиц брались значение *ipm* или ранг, часть речи (если требуется) и объем подкорпуса (количество слов). При расчете корреляции между жанровыми подкорпусами НКРЯ и веб-корпусами коэффициент корреляции и для существительных, и для глаголов оказался выше всего в публицистическом подкорпусе, что говорит о «большой публицистичности» веб-корпусов [Захаров, 2015б, с. 226].

Таким образом, напрашивается вывод о том, что при работе с веб-корпусами, несмотря на их кажущуюся репрезентативность, нужно задумываться о верификации и достоверности получаемых данных.

### **13.3. Применение статистических методов в корпусных исследованиях**

Среди методов, применяемых в корпусной лингвистике и положительно зарекомендовавших себя в разнообразных лингвистических исследованиях, следует отметить такие важнейшие статистические методы, как сводка и группировка материалов статистического наблюдения, абсолютные и относительные статистические величины, корреляционный и регрессионный анализ, дисперсионный и факторный анализ, хи-квадрат тест, меры ассоциации — Mutual Information (MI), log-likelihood, t-score, log-Dice, c-value. Ученые отмечают, что статистический аппарат, применяемый в корпусах текстов, позволяет пользователям ранжировать результаты поиска по разным параметрам и задавать пороговые значения, что приводит к выдаче наиболее значимой информации [Хохлова, 2008]. Статистический анализ позволяет лингвистам на основе отмеченных в отдельном корпусе характерных черт сделать обобщения, соотнести между собой разные подкорпусы, оценить случайность и взаимозависимость величин и сделать выводы, касающиеся как всего языка в целом, так и отдельных подъязыков. Существует связь между использованием вычислительных и, следовательно, алгоритмических и статистических методов, с одной стороны, и качественными изменениями результатов научных наблюдений, вытекающими из данного подхода [Tognini-Bonelli, 2001].

### 13.3.1. Корпусный анализ фразеологии

В рамках статистических исследований по фразеологии ставится вопрос о применении для этих целей корпусов текстов [Баранов, Вознесенская, Добровольский и др., 2017], так как существует ряд особенностей фразеологии, затрудняющих ее корпусный анализ:

- Употребление фразеологизмов существенно зависит от типа дискурса. Это касается прежде всего частоты употребления: наиболее насыщенными в этом отношении являются публицистические тексты, где регулярно встречаются многочисленные штампы, ср.: *ловить момент, не за горами, задавать тон, прекрасная половина, блюстители порядка, на автопилоте, белая смерть*. С точки зрения разнообразия используемых единиц выделяются устная речь и наиболее близкий к ней язык драматургии, где представлены выражения всех стилистических регистров. Кроме того, для каждого типа дискурса характерны идиомы с определенной стилистической окраской: для публистики — журнализмы, советизмы, элементы языка советской идеологии (*со школьной скамьи, поднимать голову, время «Ч», горячая точка, сердце нашей родины, утечка мозгов, смена караула, верхушка айберга, враг народа, важнейшее из искусств, почтовый ящик*); для разговорной речи и драматургии — разговорные идиомы (*на коне, держать дистанцию, на птичьих правах, как миленький*), сниженные выражения (*Вася Пупкин, рога пообломать, взять за жабры*) и просторечия (*идти на попятный, хошь не хоши, чин чинарем, руки в боки*), а также жаргонизмы. Для художественной литературы более характерны книжные идиомы, некоторые высокие и устаревшие выражения (*смирение паче гордости, взять грех на душу, тяжелый крест, проливать кровь, предавать огню, душой и телом*).
- Формирование полного словарника идиом, отражающего современный узус, затруднено в силу ряда обстоятельств. В большинстве текстов либо идиомы встречаются относительно редко, либо круг этих выражений ограничен. Современные фразеологические словари не успевают за изменениями, которые происходят в языке: многие выражения, упомянутые в словарях, уже вышли из употребления, а, скажем, фразеологизмы в интернет-общении описаны недостаточно

(яростно плюсую, капитан очевидность, запастись попкорном, словить лулзов).

- Яркой особенностью фразеологии является наличие вариантов у многих словарных единиц, что делает затруднительным или практически невозможным автоматический поиск в тексте и подсчет вхождений многих выражений, например: *под крылом/крыльшком, с какого бока/боку; навешивать ярлык/ярлыки* (морфологические варианты); *воротить нос/воротить носом; оставить мокре место/мокрого места не оставлять* (синтаксические варианты); *в упор не видеть/не замечать; спустить собак/спустить полканы* (лексические варианты); *иметь/затаить зуб (на кого-либо); есть зуб (у кого-либо/на кого-либо); вырос зуб (у кого-либо/на кого-либо); по разные стороны баррикад [быть/находиться]; другая сторона баррикад; по другую сторону баррикад [быть/находиться]* (лексико-синтаксические варианты). Широкое варьирование компонентного состава и отдельных компонентов идиом затрудняет определение словарной формы. В ряде случаев неочевидно, следует ли считать два или три выражения вариантами одной леммы или разными фразеоглизмами: *какого лешего? за каким лешим? отпустить душу/душеньку [на покаяние]; отпустить на покаяние (кого-либо); товарищ по несчастью; собрат по несчастью; держать руку на пульсе; держать руку на рычаге/рычагах; удержка/удержу не знать; удержу нет/не стало (на кого-либо); без удержану.*
- Наконец, последняя группа затруднений, имеющая самое непосредственное отношение к определению частотности, создается значимыми индивидуальными различиями в употреблении фразеоглизмов у разных авторов. Эти различия касаются как набора частотных выражений идиолекта, так и нежелания конкретного автора использовать идиомы.

Для создания полного представления о частотности идиом в языке в целом возникла идея создать фразеологически ориентированные корпусы, исключив из рассмотрения авторов, не употребляющих идиомы. Для решения этой задачи тексты некоторых авторов, представленные в корпусе русской прозы [Баранов, Вознесенская, Добровольский и др., 2017], были проанализированы с точки зрения частотности употребления фразеологии. На основе полученных

данных для каждого автора была выведена средняя относительная частота употребления идиом по выбранным произведениям. Результаты представлены в табл. 13.4.

Таблица 13.4. Идиоматичность в текстах некоторых авторов  
(на 1 тыс. словаупотреблений)

Авторы	Средняя частота употребления идиом
Е. Гинзбург	14
Юз Алешковский	11,1
В. Шендерович	7,6
Ю. Трифонов	5,3
Абрам Терц	5,2
Э. Рязанов	5,2
А. и Б. Стругацкие	5,1
В. Войнович	5
Саша Соколов	2,9
В. Астафьев	2,4
А. Приставкин	1,9
В. Богомолов	1
В. Распутин	0,6

В табл. 13.4 присутствуют как высокочастотные авторы, так и низкочастотные. Авторы исследования считают «3» пороговым значением частоты использования идиом, позволяющим включить автора во фразеологически ориентированный корпус. Для ряда писателей, которые по этому критерию должны быть отнесены к числу «неидиоматичных», такая характеристика представляется сомнительной (например, для Саши Соколова). При сопоставлении частоты употребления фразеологизмов в разных произведениях одного автора обнаруживается достаточно большое варьирование (табл. 13.5).

Из табл. 13.5 видно, что варьирование по произведениям тем выше, чем более «фразеологичен» автор, и наоборот, варьирование по произведениям тем ниже, чем менее «фразеологичен» автор. Кроме внешнего варьирования (различной частоты использования иди-

ом в разных произведениях одного автора) тексты произведений часто демонстрируют и внутреннее варьирование (различную частоту употребления идиом внутри одного произведения). Так, внутреннее варьирование в повести А. и Б. Стругацких «Понедельник начинается в субботу» выглядит следующим образом:  $9 + 10 + 17 + 9 + 23 + 8 + 9 + 32 + 20$ , где цифры обозначают абсолютную частоту из расчета на каждый последовательный фрагмент в 40 тыс. знаков.

*Таблица 13.5. Частота употребления фразеологизмов в разных произведениях одного автора*

<i>Юз Алешковский</i>	
Кенгуру	8,3
Маскировка	11,7
Николай Николаевич	16,6
Синенький скромный платочек	14
Среднее	11,1
<i>Саша Соколов</i>	
Между собакой и волком	6,5
Палисандрья	1,75
Школа для дураков	2,3
Среднее	2,9
<i>Валентин Распутин</i>	
Деньги для Марии	0,4
Живи и помни	0,5
Пожар	0,36
Последний срок	0,46
Прощание с Матерой	0,9
Среднее	0,6

Приведенный анализ показывает, что методика оценки идиomaticности авторов по произвольно выбранному фрагменту размером 40 тыс. знаков не работает. Необходимо исследовать частоту употребления идиом во всех произведениях автора, что по очевид-

ным причинам весьма затруднительно. Кроме того, создание фразеологически ориентированных корпусов публистики практически невозможно в силу того, что в состав этих корпусов входят тексты, написанные сотнями, если не тысячами авторов, каждый из которых обладает индивидуальными особенностями употребления идиом. Вопрос о создании фразеологически ориентированных корпусов русской прозы, детективов и драматургии остается открытым. Вероятно, более адекватную оценку идиоматичности авторов можно получить, используя в качестве «мерила» базовый словарь из 300–400 идиом, достаточно частотных для литературного языка и отражающих основные семантические поля русской фразеологии. Таким образом, создание фразеологически ориентированных корпусов превратилось из вспомогательной процедуры, предваряющей составление Частотного словаря, в самостоятельную нетривиальную задачу, которая, возможно, будет решена в процессе создания этого словаря [Баранов, Вознесенская, Добровольский и др., 2017].

### ***13.3.2. Диахронические исследования грамматики***

Использование корпуса Google Books Ngram для изучения аспектуальной системы русского языка продемонстрировали В. Д. Соловьев, В. В. Бочкарев и Л. А. Янда [2017]. Данный корпус позволяет анализировать и сопоставлять динамику частот словоупотреблений для изучения эволюции как отдельных слов, так и всего лексикона языка в целом, а также культурных трендов в обществе. Он снабжен удобной визуализацией в виде графиков частот (<https://books.google.com/ngrams/>). Кроме собственно частот словоупотреблений весьма информативной является форма графиков. Естественным является предположение, что частоты употребления семантически идентичных слов (например, словоизменение внутри одной леммы: *читать — читал*) под влиянием внешних факторов меняются схожим образом, то есть графики частот имеют схожую форму.

Указанные авторы исследовали аспектуальную систему русского языка, которая находится в процессе становления, многие ее элементы получили неоднозначное освещение в литературе и вызвали большие споры. Неясным остается характер суффиксального образования вторичных имперфективов и префиксального образования перфективов с точки зрения словообразовательного или словоизменительного характера этих процессов.

Корреляции частот словоупотребления сравниваются внутри следующих групп слов: глаголов: 1) случайно выбранные (в качестве *baseline*); 2) словоизменительные формы внутри одной леммы (*читать* — *читал*); 3) аспектуальные пары с префиксальным образованием перфективов (*делать* — *сделать*); 4) аспектуальные пары с суффиксальным образованием вторичных имперфективов (*убаюкать* — *убаюкивать*). Словоизменительная парадигма была взята по открытому ресурсу OpenCorpora (<http://opencorpora.org/>), основанному на словаре А. А. Зализняка. Случайная выборка включала 20 тыс. пар словоформ из парадигм 6947 глаголов, все формы полной парадигмы которых присутствуют в Google Books Ngram. Рассматриваются частоты словоупотребления с 1920 по 2005 г., используется корреляция по Пирсону.

В результате исследования для случайно выбранных пар глаголов и пар словоизменительных форм внутри леммы коэффициенты корреляции по Пирсону оказались равны 0,0404 и 0,1864 соответственно. Для случайно выбранных пар корреляции быть не должно, результат ожидаем. Для пар словоизменительных форм корреляция оказалась мала (что несколько странно и требует дальнейших исследований), но статистически значима ввиду огромного числа проанализированных данных. Для префиксальных и суффиксальных аспектуальных пар глаголов коэффициенты корреляции равны 0,3020 и 0,2435 соответственно. Это соотношение также неожиданно. Считается, что суффиксальное образование вторичных имперфективов — это словоизменение, а префиксальное образование перфективов — словообразование. Если это так, то это должно приводить к большей семантической схожести и более высокому коэффициенту корреляции перфективов с вторичными имперфективами. Полученный результат можно трактовать как аргумент в пользу того, что с точки зрения этого противопоставления оба способа образования аспектуальных пар имеют примерно одинаковый статус, префиксальные и суффиксальные пары ведут себя одинаково, нет статистически значимой разницы между их грамматическими профилями.

Вероятно, одним из влияющих факторов является то, что при образовании вторичного имперфектива часто привносится дополнительное значение многократности действия (*пить* —  *выпить* —  *выпивавть*). Другим фактором является выборка глаголов: в единственной существующей аспектуальной базе для русского языка

[<http://emptyprefixes.uit.no>] содержит только тройки «базовый имперфектив — естественный перфектив — вторичный имперфектив». Эта база данных не включает пары, в которых вторичный имперфектив образуется из специализированного перфектива. Учет таких пар может привести к увеличению соответствующего коэффициента корреляции.

В целом не очень высокие значения коэффициентов корреляции авторы объясняют многозначностью слов. В настоящее время нет хорошего способа автоматически выделять различные значения слов и получать для них различные графики. Для нивелирования трудностей с многозначностью были выделены 22 глагола, не имеющие абсолютно различных значений (*делать, работать, казаться, играть, просить, верить, ставить, звать, звонить, хранить, рисовать, прятать, влечь, плевать, жечь, расстить, мерзнуть, жрать, копать, нюхать, красить, щупать*). Рассмотрен более узкий временной интервал — с 1950 г., с устоявшимся современным русским языком и большим количеством изданных книг по сравнению с предшествующим периодом. Для выбранных 22 глаголов среднее значение корреляции внутри словоизменения оказалось равно 0,568, а корреляция между базовым имперфективом и естественным перфективом — 0,758. Далее для этих слов были выбраны некоторые специализированные перфективы (*переделать* и т. п.), подсчитаны коэффициенты корреляции между ними и соответствующими вторичными имперфективами. Среднее значение оказалось равно 0,633. Таким образом, и после устранения явной многозначности и учета специализированных перфектипов все равно корреляция между базовым имперфективом и естественным оказалась выше. Этот результат указывает на то, что префиксальный способ образования естественного перфектива сохраняет значение исходного слова в не меньшей степени, чем суффиксальное образование вторичных имперфективов [Соловьев, Бочкарев, Янда, 2017].

#### 13.4. 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,

особенностях их сочетаемости, управления и т. п. 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 будет содержать самые 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*,  $\log\text{-}likelihood$ , которые чаще всего используются при вычислении степени близости между компонентами словосочетаний в корпусе. Однако ряд исследований показывает, что это всего лишь самые популярные меры, в то время как более эффективными оказываются другие, такие как  $\log\text{-}Dice$ ,  $MI.\log f_{min}$ , *sensitivity* [Evert, Krenn, 2001; Pecina, 2009; Zakharov, 2017].

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 [Хохлова, 2010]. Исследование осуществлялось с помощью корпуса-менеджера CQP (<http://corpus1.leeds.ac.uk/ruscorpora.html>) на базе корпуса русских газетных текстов 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 [Шаров, 2003] были отобраны существительные, входящие в первую тысячу самых частотных слов. Далее по Малому академическому словарю (МАС) (1981–1984) проверялось, имеют ли данные слова омонимы, которые могли бы исказить 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 [Борисова, 1995]. В случае отсутствия словарных статей для данного слова или ограниченной информации о его сочетаемости, представленной в словаре, такое слово тоже исключалось.

чалось 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*. В табл. 13.6 приведены левые коллокаты 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, то есть 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 — обладают довольно высокими показателями меры *MI*.

Таблица 13.6. 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

Исследование показало, что в диапазоне значений меры  $MI$  от 0 до 1 не были найдены словосочетания, которые можно было бы причислить к устойчивым. 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 общязыковых корпусов, в ряде случаев можно увидеть существенные отличия в сочетаемости, отражающие особенности авторского словаупотребления. Таким образом, можно утверждать, что описанные выше методы и средства могут быть эффективно использованы для изучения и создания словарей языка писателей, для выявления особенностей сочетаемости в рамках того или иного стиля или хронологического периода [Хохлова, 2010].

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

### **Вопросы и задания для самоконтроля**

1. Дайте определение следующим понятиям: *нормированная частота, регистр, коллокат, коллокация*.
2. Какие лингвистические особенности регистра художественной прозы можно выявить с помощью корпусной методологии?
3. Какие перспективы открывают перед пользователями синтаксические корпусы?
4. Какие перспективы открывают перед пользователями морфологические корпусы?
5. Назовите особенности метаразметки диалектного корпуса текстов.
6. Перечислите методы корпусной лингвистики.
7. С какими проблемами в области аннотирования устных текстов сталкиваются составители корпуса ОРД и Учебного прагматического корпуса?
8. Как можно оценить «идиоматичность» в текстах корпусными методами?
9. В каких исследованиях можно применить корреляцию частот словоупотребления?
10. Предложите варианты использования корпусных методов в преподавании иностранного языка.
11. Назовите перспективные, на ваш взгляд, направления развития корпусной лингвистики.

## Заключение

Корпусная лингвистика представляет собой новое направление в лингвистической науке, позволяющее проводить исследование единиц любого языкового уровня в реальном их употреблении, то есть 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, то 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 и, возможно, уже устарела, что на смену тем или иным конкретным программам придут более совершенные и многофункциональные, появятся новые корпусы, изменятся их адреса в Интернете. Свою задачу авторы видели в том, чтобы дать описание нового направления в лингвистике как такого, дать примеры корпусных исследований разного типа и охарактеризовать состояние корпусной лингвистики в конце второго десятилетия XXI в.

Можно сказать, корпусная лингвистика быстро завоевала центральные позиции в языкоznании. Куда же плыть дальше? Некоторые направления кажутся очевидными.

Во-первых, это дальнейшее взаимовлияние и взаимопроникновение корпусной и компьютерной лингвистики, они нуждаются друг в друге.

---

#### Часть 4. Заключение

---

Во-вторых, будет наблюдаться все более широкое распространение корпусной лингвистики — как материала и как метода — на всю сферу гуманитарных исследований — историю, социологию, литературоведение и т. д. Уже сегодня на базе корпусной методологии фактически сформировалась новая наука — культурометрия (*culturomics*) (<http://www.culturomics.org/>).

Но это еще не все. Текстоориентированный социальный и культурный опыт человечества в виде корпусов текстов (в широком смысле) получает инструмент, позволяющий надеяться, что мы научимся автоматически извлекать из текстов знание. Представляется, что корпусная лингвистика вместе с психолингвистикой и нейролингвистикой сформируют новую науку — интегрированную эмпирическую лингвистику, которая позволит глубже, чем до сих пор, понять фундаментальную природу языка.

# Темы докладов, рефератов, курсовых работ

1. Способы использования корпусов в лингвистических исследованиях.
2. Способы использования корпусов в лексикографии.
3. Изучение средств обработки корпусных данных, представленных на языке XML.
4. Исследование механизмов взаимодействия корпуса текстов и электронной картотеки (корпусы цитат).
5. Токенизация текстов.
6. Унификация текстов внутри корпуса.
7. Автоматическая морфологическая разметка текстов разных периодов.
8. Исследование наборов метаданных.
9. Методы снятия морфологической неоднозначности.
10. Анализ функций сегментных внеалфавитных графем (межморфемный дефис, межслоговой дефис, межсловный дефис, апостроф).
11. Проблема строчных и прописных букв в корпусах текстов (имена собственные и нарицательные, сплошная и начальная капитализация).
12. Проблема омографии — акцентно-ориентированный морфологический анализ.
13. Разработка модуля преобразования каллиграфем (выделение полужирным, курсивом, подчеркивание) в теги языка XML.
14. Анализ функций точки (и других знаков препинания) с точки зрения структурной разметки текста.
15. Методы выделения структурных элементов текста.

16. Составные лексемы.
17. Коллокации.
18. Коорпусы текстов в проекте TEI.
19. Стандарты EAGLES (обзор).
20. Форматы CDIF и XCES (обзор).
21. Анализ и описание корпусного менеджера Xaira.
22. Анализ и описание корпусного менеджера NoSketch Engine.
23. Анализ и описание корпусного менеджера CQP.
24. Анализ и описание интерфейса WebCorp.
25. Анализ и описание интерфейса системы Corpus.Byu.Edu.
26. Анализ и описание корпусов системы Corpus.Byu.Edu.
27. Анализ и описание функциональных возможностей системы Corpus.Byu.Edu.
28. Сравнительный анализ возможностей разных корпусов.
29. Использование корпусов в социологии и социолингвистике.
30. Использование корпусов в этнолингвистике.
31. Семантическая разметка корпуса текстов русского языка.
32. Разработка редактора размеченных текстов.
33. Статобработка размеченных текстов.
34. Создание веб-сайта по корпусной лингвистике.
35. Обработка нестандартных элементов текстов при загрузке в корпус.
36. Снятие морфологической неоднозначности при разметке.
37. Сравнение морфологических анализаторов.
38. Автоматическое выявление терминов на базе корпусов.
39. Автоматическое выявление терминологических сочетаний.
40. Семантическая разметка корпуса текстов русского языка.
41. Разработка редактора размеченных текстов.
42. Статобработка размеченных текстов.
43. Создание веб-сайта по корпусной лингвистике
44. Обработка нестандартных элементов текстов при загрузке в корпус.
45. Снятие морфологической неоднозначности.
46. Сравнение морфологических теггеров.
47. Автоматическое выявление терминов.

48. Автоматическое выявление терминологических сочетаний.
49. Статистика предложных сочетаний (по корпусным данным) (на уровне лексических единиц).
50. Статистика предложных сочетаний (по корпусным данным) (на уровне синтаксических моделей).
51. Статистический анализ синтаксического корпуса русского языка (по синтаксическим отношениям и типам текстов).
52. Оценка автоматических методов выявления словосочетаний (различные меры ассоциации).
53. Оценка автоматических методов выявления словосочетаний (с точки зрения различных функциональных стилей).
54. Оценка автоматических методов выявления словосочетаний (учет и влияние величины контекста).
55. Выявление устойчивых сочетаний для словаря общеупотребительной специальной лексики.
56. Автоматизированные процедуры графематического анализа при создании корпусов.
57. Создание параллельных корпусов.
58. Исследования на параллельных корпусах.
59. Создание параллельного чешско-русского корпуса.
60. Словарь (корпус) русско-чешских устойчивых сочетаний.
61. Создание специальных корпусов.
62. Исследования на специальных корпусах.
63. Выделение именованных сущностей в морфологически размеченном тексте.
64. Разработка формата хранения синтаксических корпусов с сохранением структурной неоднозначности.
65. Особенности выделения разрывных генитивных именных групп на корпусном материале.
66. Автоматизированные процедуры: подготовка текстов для корпусов.
67. Создание сверхбольших корпусов методами сканирования веб-пространства.
68. Автоматическое построение лексико-семантических полей на материале корпусов.

---

Темы докладов, рефератов, курсовых работ

---

69. Корпусно-ориентированные методы построения словарей и терминосистем.
70. Сравнение региональных вариантов одного и того же языка (русский, английский).
71. Сравнение и оценка корпусных менеджеров.
72. Сравнение и оценка корпусов.
73. Социолингвистические и культурометрические исследования на основе корпусов.
74. Исторические исследования на основе корпусов.
75. История корпусной лингвистики.
76. История корпусной лингвистики в СССР и в России.
77. Составление хрестоматии по корпусной лингвистике.
78. Составление глоссария по корпусной лингвистике.
79. Квантитативное описание русских предложных конструкций.
80. Корпусные исследования особенностей перевода (*translation universals*) на базе параллельных корпусов.

# Рекомендуемая литература

## Основная

- Баранов А. Н. Введение в прикладную лингвистику. М.: Эдиториал УРСС, 2001.
- Грудева Е. В. Корпусная лингвистика: учеб. пос. 3-е изд., стереотип. М.: ФЛИНТА, 2017.
- Захаров В. П. Корпусная лингвистика: учеб.-метод. пос. СПб.: Изд-во С.-Петербург. ун-та, 2005.
- Копотев М. В. Введение в корпусную лингвистику. Прага: Animedia Company, 2014.
- Baker P., McEnery T., Hardie A. A glossary of corpus linguistics. Edinburgh: Edinburgh University Press, 2006.
- Biber D., Conrad S., Reppen R. Corpus linguistics: Investigating language structure and use. Cambridge: Cambridge University Press, 1998.
- McEnery T., Hardie A. Corpus linguistics: method, theory and practice. Cambridge: Cambridge University Press, 2012.
- Xiao R., Tono Y. Corpus-based language studies: An advanced resource book. London; New York: Routledge, 2006.

## Дополнительная

- Беляева Л. Н. Лексикографический потенциал параллельного корпуса текстов // Труды международной конференции «Корпусная лингвистика — 2004». СПб.: Изд-во С.-Петерб. ун-та, 2004. С. 55–64.
- Бочаров В. В., Грановский Д. В. Программное обеспечение для коллективной работы над морфологической разметкой корпуса // Труды международной конференции «Корпусная лингвистика — 2011». СПб.: Изд-во С.-Петерб. ун-та, 2011. С. 104–109.
- Гиндин С. И. О культурных корнях корпусной лингвистики и ее возможных импликациях для теоретического и прикладного языковедения // Труды международной конференции «Корпусная лингвистика — 2015». СПб.: Изд-во С.-Петерб. ун-та, 2015. С. 170–180.

---

## Рекомендуемая литература

---

- Гришина Е. А., Савчук С. О.* Национальный корпус русского языка как инструмент для изучения вариативности грамматических норм // Труды международной конференции «Корпусная лингвистика — 2008». СПб.: Изд-во С.-Петерб. ун-та, 2008. С. 161–169.
- Захаров В. П., Масевич А. Ц.* Диахронические исследования на основе корпуса русских текстов Google Books Ngram Viewer. Структурная и прикладная лингвистика. Вып. 10. СПб.: Изд-во С.-Петерб. ун-та, 2014. С. 303–327.
- Захаров В. П.* Сочетаемость через призму корпусов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Вып. 14 (21). Т. 1. М.: Изд-во РГГУ, 2015. С. 667–682.
- Захаров В. П., Азарова И. В., Митрофанова О. А.,* и др. Модель программно-лингвистического комплекса для создания и использования специализированных корпусов русского языка. СПб.: Изд-во С.-Петерб. ун-та, 2019.
- Митрофанова О. А.* Вероятностное моделирование тематики русскоязычных корпусов текстов с использованием компьютерного инструмента GenSim // Труды международной конференции «Корпусная лингвистика — 2015». СПб.: Изд-во С.-Петерб. ун-та, 2015. С. 332–343.
- Перцов Н. В.* О роли корпусов в лингвистических исследованиях // Труды международной конференции «Корпусная лингвистика — 2006». СПб.: Изд-во С.-Петерб. ун-та; РХГА, 2006. С. 318–331.
- Савчук С. О.* Устная публичная речь в мультимедийном модуле НКРЯ // Труды международной конференции «Корпусная лингвистика — 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. С. 315–320.
- Сичинава Д. В.* Национальный корпус русского языка: очерк предыстории // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005. С. 21–30.
- Фрэнсис У. Н.* Проблемы формирования и машинного представления большого корпуса текстов // Новое в зарубежной лингвистике: Проблемы и методы лексикографии. М.: Прогресс, 1983. С. 334–335.
- Хохлова М. В.* Особенности статистических мер при выделении биграмм // Труды международной конференции «Корпусная лингвистика — 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. С. 349–354.
- Шайкевич А. Я.* Меры лексического сходства частотных словарей // Труды международной конференции «Корпусная лингвистика — 2015». СПб.: Изд-во С.-Петерб. ун-та, 2015. С. 434–442.
- Шаров С. А.* Представительный корпус русского языка в контексте мирового опыта // Научно–техническая информация. Сер. 2. № 6. С. 12–16.
- Шерстинова Т. Ю.* «Один речевой день» на временной шкале: о перспективах исследования динамических процессов на материале звукового корпуса

- // Вестник СПбГУ. Филология. Востоковедение. Журналистика. Сер. 9. Вып. 4. Ч. 2. СПб.: Изд-во С.-Петерб. ун-та, 2008. С. 227–235.
- Benko V., Butašová A.* Teaching corpus linguistics with Aranea web corpora // Труды международной конференции «Корпусная лингвистика — 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. С. 16–22.
- Church K. W., Hanks P.* Word Association Norms, Mutual Information and Lexicography // Computational Linguistics. 1990. Vol. 16, no. 1. P. 22–29.
- Contemporary Corpus Linguistics / ed. P. Baker.* London: Continuum, 2009. 357 p.
- Corpora and Language Teaching / ed. K. Aijmer.* Amsterdam; Philadelphia: John Benjamins, 2009. 232 p.
- Corpus Linguistics. An International Handbook.* Vol. 1, 2 / eds A. Lüdeling, M. Kyö. Berlin; New York: Walter de Gruyter, 2008. 1402 p.
- Fillmore Ch. J., Atkins B. T. S.* Starting Where the Dictionaries Stop: The Challenge of Corpus Lexicography // Computational Approaches to the Lexicon. Oxford: Oxford University Press, 1994. P. 349–393.
- Kilgarriff A.* Googleology is bad science // Computational Linguistics. 2007. Vol. 33, no. 1. P. 147–151.
- Kunilovskaya M., Kutuzov A.* A quantitative study of translational Russian (based on a translational learner corpus) // Труды международной конференции «Корпусная лингвистика — 2015». СПб.: Изд-во С.-Петерб. ун-та, 2015. С. 33–40.
- Leech G.* The State of the Art in Corpus Linguistics / eds K. Aijmer, B. Altenberg. English Corpus Linguistics. Studies in Honour of Jan Svartvik. London: Longman, 1991. P. 8–29.
- McEnergy T., Xiao R., Tono Y.* Corpus-Based Language Studies. An Advanced Resource Book. London: Routledge, 2006.
- Scrivner O., Trapido I., Lee J.* Text mining toolkit for digital corpora // Труды международной конференции «Корпусная лингвистика — 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. С. 72–77.
- Sinclair J.* Corpus Concordance Collocation. Oxford: Oxford University Press, 1991.
- Stefanowitsch A.* A lot of data: textually distinctive collexemes in a corpus of scientific English // Труды международной конференции «Корпусная лингвистика — 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. С. 85–95.
- Tognini-Bonelli E.* Corpus Linguistics at Work. Amsterdam: John Benjamins, 2001.
- WaCky! Working papers on the Web as Corpus / eds M. Baroni, S. Bernardini.* Bologna: Gedit, 2006.

## СПИСОК ЦИТИРУЕМЫХ ИСТОЧНИКОВ

- 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.
- Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л.* и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003–2005 (результаты и перспективы). М.: Индрик, 2005. С. 193–214.
- Асиновский А. С., Богданова Н. В., Русакова М. В.* и др. 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.: Изд-во РГГУ, 2008. С. 488–494. URL: <http://www.dialog-21.ru/media/1796/76.pdf> (дата обращения: 16.06.2019).
- Баранов А. Н.* Введение в прикладную лингвистику. М.: Эдиториал УРСС, 2001.
- Баранов А. Н.* Лингвистическая экспертиза текста. Теоретические основания и практика. М.: Флинта; Наука, 2007.
- Баранов А. Н., Плунгян В. А., Рахилина Е. В.* Путеводитель по дискурсивным словам русского языка. М.: Помовский и партнеры, 1993.
- Баранов А. Н., Вознесенская М. М., Добровольский Д. О.* Статистические исследования во фразеологии: проблема фразеологичности корпусов текстов // Труды международной конференции «Корпусная лингвистика — 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. С. 114–120.
- Беликов В., Копылов Н., Пиперски А.* и др. Корпус как язык: от масштабируемости к дифференциальной полноте // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Бекасово, 29 мая — 2 июня 2013 г.). Вып. 12 (19). М.: Изд-во РГГУ, 2013. Т. 1. С. 84–95.
- Беликов В. И., Селегей В. П., Шаров С. А.* Пролегомены к проекту Генерального интернет-корпуса русского языка (ТИКРЯ) // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной между-

- народной конференции «Диалог» (Бекасово, 30 мая — 3 июня 2012 г.). Вып. 11 (18). М.: Изд-во РГГУ, 2012. Т. 1. С. 37–49.
- Беляева Л. Н.* Лексикографический потенциал параллельного корпуса текстов // Труды международной конференции «Корпусная лингвистика — 2004». СПб.: Изд-во С.-Петерб. ун-та, 2004. С. 55–64.
- Беляева Л. Н.* Словари систем машинного перевода и параллельные корпусы текстов: проблемы корреляции // Труды международной конференции «Корпусная лингвистика — 2015». СПб.: Изд-во С.-Петерб. ун-та, 2015. С. 102–110.
- Блинова О. В.* Побудительная реплика в диалогическом окружении: пары реплик типа «императив + вербальная реакция» и способы их разметки в речевом корпусе // Труды международной конференции «Корпусная лингвистика — 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. С. 128–133.
- Богданова С. Ю.* Исследование слова и предложения компьютерными методами // Слово в предложении: кол. монография / отв. ред. Л. М. Ковалева, ред. С. Ю. Богданова, Т. И. Семенова. Иркутск: ИГЛУ, 2010. С. 194–213.
- Богданова-Бегларян Н. В.* Устная спонтанная речь: судьба некоторых грамматических единиц // Труды международной конференции «Корпусная лингвистика — 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. С. 134–139.
- Большой толковый словарь русского языка / гл. ред. С. А. Кузнецова. СПб.: Но-rint, 1998.
- Борисова Е. Г.* Слово в тексте. Словарь коллокаций (устойчивых словосочетаний) русского языка с англо-русским словарем ключевых слов. М.: Филология, 1995.
- Гарабик Р., Захаров В. П.* Параллельный русско-словацкий корпус // Труды международной конференции «Корпусная лингвистика — 2006». СПб.: Изд-во С.-Петерб. ун-та, 2006. С. 81–87.
- Гвишиани Н. Б.* Практикум по корпусной лингвистике: учеб. пос. по английскому языку. М.: Высшая школа, 2008.
- 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.
- Гришина Е. А., Савчук С. О.* Корпус звучащей русской речи в составе Национального корпуса русского языка // Компьютерная лингвистика и интеллектуальные технологии (по материалам ежегодной международной конференции «Диалог-2008»). М.: РГГУ, 2008. С. 125–132.
- Грудева Е. В.* Корпусная лингвистика: Учебное пособие. 3-е изд., стереотип. М.: Флинта, 2017.
- Ерилов А. П.* К методологии построения диалоговых систем: Феномен деловой прозы. Новосибирск: ВЦ СО АН СССР, 1979. (Препринт / АН СССР. Сиб. отд-ние. ВЦ; 156).

---

## Список цитируемых источников

---

- Зализняк А. А., Левонтина И. Б., Шмелев А. Д. Ключевые идеи русской языковой картины мира: сб. ст. М.: Языки славянской культуры, 2005.
- Засорина Л. Н. Частотный словарь русского языка. М.: Русский язык, 1977.
- 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.: Наука, 2005. С. 166–171.
- Захаров В. П. Дистрибутивно-статистический анализ как инструмент автоматизации формирования семантических полей (на примере поля «империя») // XV Международная конференция по компьютерной и когнитивной лингвистике TEL — 2018 (31 октября — 3 ноября 2018 г., Казань, Россия): сб. тр.: в 2 т. Казань: Изд.-во АН РТ, 2018. Т. 2. С. 163–180.
- Захаров В. П. Оценка качества Интернет-корпусов русского языка // Труды международной конференции «Корпусная лингвистика — 2015». СПб.: Изд-во С.-Петерб. ун-та, 2015. С. 218–229.
- Захаров В. П. Поисковые системы Интернета как инструмент лингвистических исследований // Русский язык в Интернете: сб. ст. Казань: Отечество, 2003. С. 48–59.
- Захаров В. П. Сочетаемость через призму корпусов // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог» (Москва, 27–30 мая 2015 г.). Вып. 14 (21). М.: Изд-во РГГУ, 2015. Т. 1. С. 667–682.
- Захаров В. П., Масевич А. Ц. Диахронические исследования на основе корпуса русских текстов Google Books Ngram Viewer // Структурная и прикладная лингвистика. Вып. 10. СПб.: Изд-во С.-Петерб. ун-та, 2014. С. 303–327.
- Зубов А. В., Зубова И. И. Информационные технологии в лингвистике: учеб. пос. М.: Издательский центр «Академия», 2004.
- Камишилова О. Н. Лингвистический объект в интерпретации корпусных технологий: в пользу доказательной парадигмы // Прикладная лингвистика в науке и образовании: лингвистические технологии и инновационная образовательная среда: кол. монография. СПб.: Лема, 2010. С. 84–96.
- Камишилова О. Н. Учебный корпус текстов: потенциал, состав, структура. СПб.: ООО «Книжный Дом», 2012.
- Камишилова О. Н., Захаров В. П. Корпусное исследование прилагательных в иноязычной речи на разных этапах усвоения чужого языка (Corpus-driven Research of Adjectives in Learner Language Acquisition) // Proceedings of the R. Piotrowski's Readings in Language Engineering and Applied Linguistics. Saint Petersburg, Russia, November 27, 2017. St. Petersburg, 2017. P. 177–190. URL: <http://ceur-ws.org/Vol-2233/> (дата обращения: 14.06.2019).

- Камишилова О. Н., Смульская Е. Д.* Планирование лонгитюдного исследования: большие проблемы маленького корпуса // Труды VIII Международной научной конференции «Прикладная лингвистика в науке и образовании» 24–26 ноября 2016 г., Санкт-Петербург. СПб.: ООО «Книжный Дом», 2016. С. 170–175.
- Карлсон Ф.* Ранняя генеративная лингвистика и эмпирическая методология // Когнитивные категории в синтаксисе: кол. монография. Иркутск: ИГЛУ, 2009. С. 215–247.
- Копотев М. В.* Введение в корпусную лингвистику. Прага: Animedia Company, 2014.
- 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.
- Крейдлин Г. Е.* Невербальная семиотика. М.: Новое литературное обозрение, 2002.
- Крылов С. А., Фролова О. Е.* О корпусе официально-деловых текстов русского языка // Труды международной конференции «Корпусная лингвистика — 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. С. 226–230.
- Крючкова О. Ю., Гольдин В. Е.* Параметры обработки текстов для русского диалектного корпуса // Труды международной конференции «Корпусная лингвистика — 2015». СПб.: Изд-во С.-Петерб. ун-та, 2015. С. 307–314.
- Крючкова О. Ю., Гольдин В. Е.* Диалектный текстовый корпус: проблемы репрезентативности, сбалансированности, единиц хранения и выдачи // Труды международной конференции «Корпусная лингвистика — 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. С. 231–235.
- Кучаков Р., Савельев Д.* Сложность правовых актов в России: Лексическое и синтаксическое качество текстов / отв. ред. Д. Скугаревский. СПб.: ИПП ЕУСПб, 2018 (Сер. «Аналитические записки по проблемам правоприменения»). URL: [http://enforce.spb.ru/images/analit\\_zapiski/memo\\_readability\\_2018\\_web.pdf](http://enforce.spb.ru/images/analit_zapiski/memo_readability_2018_web.pdf) (дата обращения: 14.06.2019).
- Лингвистический энциклопедический словарь (ЛЭС) / отв. ред. В. Н. Ярцева. М.: Сов. энциклопедия, 1990.
- Ляшевская О. Н., Шаров С. А.* Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.
- Малый академический словарь / отв. ред. А. П. Евгеньева. 2-е изд., испр. и доп. М.: Русский язык, 1981–1984.
- 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» / отв. ред. А. Е. Кибрик. М., 2008. С. 352–355.

---

## Список цитируемых источников

---

- Машинный фонд русского языка: идеи и суждения / отв. ред. В. М. Андрющенко. М.: Наука, 1989.
- 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.
- Митрофанова О. А., Захаров В. П.* Автоматизированный анализ терминологии в русскоязычном корпусе текстов по корпусной лингвистике // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). М.: РГГУ, 2009. С. 321–328.
- Модель программно-лингвистического комплекса для создания и использования специализированных корпусов русского языка / отв. ред. В. П. Захаров, ред. И. В. Азарова, О. А. Митрофанова, А. М. Попов и др.. СПб.: Изд-во С.-Петерб. ун-та, 2019.
- Национальный корпус русского языка. URL: <http://ruscorpora.ru> (дата обращения: 16.06.2019).
- Николаева Ю. В.* Кинетические признаки структуры устного нарратива (корпусное исследование) // Проблемы компьютерной лингвистики: сб. науч. тр. / отв. ред. А. А. Кретов. Вып. 4. Воронеж: Изд-во ВГУ, 2010. С. 193–200.
- 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.
- Пиперски А. Ч.* Генеральный интернет-корпус русского языка и понятие репрезентативности в корпусной лингвистике // Современные проблемы науки и образования. 2013. № 5. URL: <http://www.science-education.ru/ru/article/view?id=9895> (дата обращения: 16.06.2019).
- Потапова Р. К.* Новые информационные технологии и лингвистика: учеб. пос. М.: Едиториал УРСС, 2005.
- Пужаева С. Ю., Зевахина Н. А., Джакупова С. С.* Контаминация конструкций в речи нестандартных русскоговорящих на материале Корпуса русских учебных текстов // Труды международной конференции «Корпусная лингвистика — 2015». СПб.: Изд-во С.-Петерб. ун-та, 2015. С. 390–397.
- Риехакайнен Е. И.* Корпус транскрибированных русских устных текстов: текущие возможности и перспективы // Труды международной конференции «Корпусная лингвистика — 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. С. 304–308.

- Рыков В. В. Корпус текстов как реализация объектно-ориентированной парадигмы // Труды Международного семинара «Диалог-2002». М.: Наука, 2002. С. 124–129.
- Савельев Д. А. О создании и перспективах использования корпуса текстов российских правовых актов как набора открытых данных // Право. Журнал Высшей школы экономики. 2018. №. 1. С. 26–44. URL: <https://law-journal.hse.ru/2018--1/218418304.html> (дата обращения: 14.06.2019).
- Сичинава Д. В. Национальный корпус русского языка: Очерк предыстории // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005. С. 21–30.
- Смит Б. Методы и алгоритмы вычислений на строках (regexp) = Computing Patterns in Strings. М.: Вильямс, 2006.
- Соловьев В. Д., Бочкарев В. В., Янда Л. А. Динамика частот употребления семантически близких слов // Труды международной конференции «Корпусная лингвистика — 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. С. 325–329.
- Тимофеева М. К. Корпусная pragматика: о возможности создания учебного корпуса // Труды международной конференции «Корпусная лингвистика — 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. С. 343–348.
- Толпегин П. В. Автоматическое разрешение кореференции местоимений третьего лица русскоязычных текстов: автореф. дис. ... канд. техн. наук. М., 2008.
- Труб В. М. О возможном подходе к семантическому описанию частей тела // Московский лингвистический журнал. 2006. Т. 8. № 1. С. 67–73.
- Фридл Дж. Регулярные выражения = Mastering Regular Expressions. СПб.: Питер, 2001.
- Хохлова М. В. Экспериментальная проверка методов выделения коллокаций // Slavica Helsingiensia 34. Инструментарий русистики: Корпусные подходы. Helsinki: Helsinki University Press, 2008. С. 343–357.
- Хохлова М. В. Исследование лексико-сintаксической сочетаемости в русском языке с помощью статистических методов на базе корпусов текстов: автореф. дис. ... канд. филол. наук. СПб., 2010.
- Шаров С. А. Частотный словарь русского языка. 2002. URL: <http://www.artint.ru/projects/frqlist.asp> (дата обращения: 14.06.2019).
- Шаров С. А., Беликов В. И., Копылов Н. Ю. и др. Корпус с автоматически снятой морфологической неоднозначностью: к методике лингвистических исследований // Компьютерная лингвистика и интеллектуальные технологии. 2015. Т. 14, № 1. С. 84–95.
- Шерстинова Т. Ю. «Один речевой день» на временной шкале: о перспективах исследования динамических процессов на материале звукового корпуса // Вестник СПбГУ. Сер. 9. Филология. Востоковедение. Журналистика. Вып. 4. Ч. 2. СПб.: Изд-во С.-Петерб. ун-та, 2008. С. 227–235.

- Шерстинова Т.Ю.* Подходы к тематическому аннотированию звукозаписей повседневного бытового общения в корпусе «Один речевой день» // Труды международной конференции «Корпусная лингвистика — 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. С. 367–372.
- Шерстинова Т.Ю.* Прагматическое аннотирование коммуникативных единиц в корпусе ОРД: микроэпизоды и речевые акты // Труды международной конференции «Корпусная лингвистика — 2015». СПб.: Изд-во С.-Петерб. ун-та, 2015. С. 451–459.
- Шерстинова Т.Ю., Рыко А.И., Степанова С.Б.* Система аннотирования в звуковом корпусе русского языка «Один речевой день» // Формальные методы анализа речи. Материалы XXXVIII Международной филологической конференции (16–20 марта 2009 г.). СПб.: Факультет филологии и искусств СПбГУ, 2009. С. 66–75.
- Щипицина Л.Ю.* Информационные технологии в лингвистике: учеб. посес. 2-е изд., стереотип. М.: Флинта; Наука, 2015.
- Эйсмонт П.М.* «КОНДУИТ»: Корпус устных детских текстов // Труды международной конференции «Корпусная лингвистика — 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. С. 373–377.
- Anthony L.* AntConc (Version 3.5.8) [Computer Software]. Tokyo: Waseda University, 2019. URL: <http://www.laurenceanthony.net/software/antconc/> (дата обращения: 14.06.2019).
- Archer D., Wilson A., Rayson P.* Introduction to the USAS category system. Benedict project report, October 2002. URL: [http://ucrel.lancs.ac.uk/usas/usas\\_guide.pdf](http://ucrel.lancs.ac.uk/usas/usas_guide.pdf) (accessed on 23.06.2019)
- Baker P., McEnery T., Hardie A.* A glossary of corpus linguistics. Edinburgh: Edinburgh University Press, 2006.
- Barlow M.* MonoConc Pro (MP 2.2) [Computer Software]. LINGUIST List 11.1411 (<https://linguistlist.org/issues/11/11-1411.html>), 2000. URL: <http://www.athel.com/mono.html> (accessed on 14.06.2019).
- Baroni M., Bernardini S.* BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004. Lisbon: ELDA, 2004. P. 1313–1316.
- Baroni M., Bernardini, S., Ferraresi A. et al.* The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora // Language Resources and Evaluation. 2009. Vol. 43, no. 3. P. 209–226.
- Beliaeva L.* Applied Lexicography and Scientific Text Corpora // Translations on Business and Engineering Intelligent Applications / eds G. Setlak, K. Markov. Rzeszow: ITHEA, 2014. P. 55–63.
- Benko V.* Aranea: Yet Another Family of (Comparable) Web Corpora / eds P. Sojka, A. Horák, I. Kopeček et al. // Text, Speech and Dialogue. 17<sup>th</sup> International Conference, TSD-2014, Proceedings. LNCS 8655. Springer International Publishing Switzerland, 2014. P. 257–264.

- Benko V., Zakharov V.P.* Very Large Russian Corpora: New Opportunities and New Challenges // Computational Linguistics and Intellectual Technologies. Vol. 15. Moscow, 2016. P.79–93.
- Biber D., Conrad S., Reppen R.* Corpus Linguistics. Investigating language structure and use. Cambridge: Cambridge University Press, 1998.
- Český národní korpus — úvod a příručka uživatele FF UK / eds J. Kocek, M. Kopřivová, K. Kučera. Praha: ÚČNK, 2000.
- Corpus Linguistics. An International Handbook / eds A. Lüdeling, M. Kytö. Vol. 1, 2. Berlin; New York: Walter de Gruyter, 2008.
- Evert S., Krenn B.* Methods for the Qualitative Evaluation of Lexical Association Measures // ACL Proceedings of 39th Annual Meeting (39<sup>th</sup> ACL & 10<sup>th</sup> EACL). Toulouse, 2001. P.188–195.
- Firth, J. R.* A synopsis of linguistic theory, 1930–1955 // Studies in Linguistic Analysis, special volume, Philological Society. Oxford: Blackwell, 1957. P.1–32. Reprinted ed. F.R. Palmer. Selected Papers of J. R. Firth, 1952–1959. London: Longman, 1968. P.168–205.
- Finegan E.* Language: its structure and use. New York: Harcourt Brace College Publishers, 2004.
- Francis W.N.* Language Corpora B.C.// Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4–8 August 1991 / ed. J. Svartvik. Berlin; New York: Mouton de Gruyter, 1992. P.17–32.
- Genres on the Web: Computational Models and Empirical Studies (Text, Speech and Language Technology) / eds A. Mehler, S. Sharoff, M. Santini. Springer Verlag, 2010.
- Hajič J.* Disambiguation of Rich Inflection: Computational Morphology of Czech. Prague: Karolinum Press, 2004.
- Hardie A.* CQPweb — combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics. 2012. No. 17 (3). P. 380–409. URL: <http://cwb.sourceforge.net/cqpweb.php> (accessed on 14.06.2019).
- Herman O., Kovář V.* Methods for Detection of Word Usage over Time // VII Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2013. Brno: Tribun EU, 2013. P.79–85.
- How to use corpora in language teaching / ed. J. M. Sinclair. Amsterdam: John Benjamins, 2004.
- Jakubíček M., Kilgarriff A., McCarthy D. et al.* Fast Syntactic Searching in Very Large Corpora for Many Languages // PACLIC 24: Proceedings of the 24<sup>th</sup> Pacific Asia Conference on Language, Information and Computation. 2010. P.741–747.
- Jakubíček M., Kilgarriff A., Kovář V. et al.* The TenTen Corpus Family // Proceedings of the 7<sup>th</sup> International Corpus Linguistics Conference. Lancaster: UCREK, 2013. P.125–127.

- Johansson S.* Some aspects of the development of corpus linguistics in the 1970s and 1980s // *Corpus Linguistics. An International Handbook* / eds A. Lüdeling, M. Kytö. Vol. 1. Berlin; New York: Walter de Gruyter, 2008. P. 33–53.
- Ide N., Romary L.* International standard for a linguistic annotation framework // *Natural language engineering*. 2003. Vol. 10, no. 3–4. P. 211–225.
- Ide N., Romary L.* Standards for language resources. Proc. of Language Resources and Evaluation Conference (LREC02). Las Palmas: The University of Las Palmas de Gran Canaria, 2002. P. 59–65.
- Kilgarriff A.* Web as corpus // Proc. of Corpus Linguistics 2001 conference / Lancaster University. Lancaster: UCREL, 2001. P. 342–344.
- Kilgarriff A., Baisa V., Bušta J.* et al. The Sketch Engine: Ten Years On // Lexicography ASIALEX, 2014a. Vol. 1. P. 7–36. URL: <https://link.springer.com/article/10.1007/s40607-014-0009-9> (accessed on 14.06.2019).
- Kilgarriff A., Grefenstette G.* Introduction to the Special Issue on Web as Corpus // *Computational Linguistics*. 2003. Vol. 29, no. 3). P. 333–347.
- Kilgarriff A., Jakubíček M., Kovář V.* et al. Finding Terms in Corpora for Many Languages with the Sketch Engine // Proceedings of the Demonstrations at the 14th Conference the European Chapter of the Association for Computational Linguistics. Sweden Gothenburg: Association for Computational Linguistics. P. 53–56. URL: [https://www.sketchengine.co.uk/wp-content/uploads/Finding\\_Terms\\_2014.pdf](https://www.sketchengine.co.uk/wp-content/uploads/Finding_Terms_2014.pdf) (accessed on 14.06.2019).
- Kilgarriff A., Rychlý P.* An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments) // Proceedings of the 45<sup>th</sup> Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. / ed. S. Ananiadou. Prague: Association for Computational Linguistics, 2007. P. 41–44.
- Kilgarriff A., Rychlý P., Jakubíček M., Rundell M.* et al. SketchEngine [Computer Software]. URL: <http://www.sketchengine.co.uk> (accessed on 14.06.2019).
- Lakoff G.* Pronominalization, Negation, and the Analysis of Adverbs // Readings in English transformational grammar / eds R. Jacobs, P. Rosenbaum. Waltham: Ginn & Co, 1970. P. 145–165.
- Leech G.* The Distribution and Function of Vocatives in American and British English Conversation // Out of Corpora. Studies in Honour of Stig Johansson / eds H. Hasselgård, S. Oksefjell. Amsterdam: Rodopi, 1999. P. 107–120.
- Ljubešić N., Erjavec T.* hrWaC and slWaC: Compiling Web Corpora for Croatian and Slovene. Text, Speech and Dialogue — 2011. Lecture Notes in Computer Science. город Springer, 2011. P. 395–402.
- Ljubešić N., Klubička F.* {bs,hr,sr}WaC — Web corpora of Bosnian, Croatian and Serbian // Proceedings of the 9th Web as Corpus Workshop (WaC-9). Gothenburg: Association for Computational Linguistics, 2014. P. 29–35.

- Lönnqvist L.* Chastotnyi slovar' sovremennoj russkoj jazyka. Uppsala: Studia Slavica Upsaliensia, 1993.
- Lüdeling A., Kyö M.* Corpus linguistics: an international handbook. Vol. 1. Berlin; New York: W. de Gruyter, 2008.
- McEnery T., Hardie A.* Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press, 2012.
- McEnery T., Wilson A.* Corpus Linguistics. Edinburgh: Edinburgh University Press, 2001.
- McWhinney B.* The CHILDES Project: Tools for Analyzing Talk. Inc. 3<sup>rd</sup> ed. Mahwah; New York: Lawrence Erlbaum Associates, 2000. Vol. 1.
- Meyer Ch. F.* English Corpus Linguistics: An Introduction. Cambridge: Cambridge University Press, 2002.
- Michel J. B.* Quantitative Analysis of Culture Using Millions of Digitized Books science // Science. 2011. No. 331. P. 176–182.
- 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. Levická, R. Garabik. Brno: Tribun, 2009. P. 249–255.
- P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.4.0. Last updated on 23<sup>rd</sup> July 2018, revision 1fa0b54. URL: <http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>. (accessed on 29.10.2018). URL: <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/> (accessed on 29.10.2018).
- Pecina P.* Lexical Association Measures. Collocation Extraction. Praha: Ústav formální a aplikované lingvistiky, 2009.
- 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. 1968. No. 3. P. 1–239.
- Postal P. M.* On the Surface Verb ‘remind’ // Linguistic Inquiry. 1970. Vol. 1. P. 37–120.
- Rayson P.* Wmatrix: a web-based corpus processing environment [Computer Software]. Computing Department, Lancaster University, 2013. URL: <http://ucrel.lancs.ac.uk/wmatrix/> (accessed on 14.06.2019).
- Renouf A., Kehoe A., Banerjee J.* WebCorp: an integrated system for web text search // Language and Computers. 2006. Vol. 59, no. 1. P. 47–67. URL: [http://rdubs.bcu.ac.uk/publ/WebCorp\\_integrated\\_system\\_DRAFT.pdf](http://rdubs.bcu.ac.uk/publ/WebCorp_integrated_system_DRAFT.pdf) (accessed on 14.06.2019).
- Rychlý P.* A lexicographer-friendly association score // Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN. Brno: Masaryk University, 2008. P. 6–9.

- Rychlý P. Manatee/Bonito — A Modular Corpus Manager // 1<sup>st</sup> Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University, 2007. P.65–70.
- Sedláček R., Smrž P. A new Czech morphological analyser ajka. In Proceedings of the 4<sup>th</sup> International Conference TSD. 2001. LNCS 2166, Springer-Verlag. 2001. P. 100–107.
- Schäfer R. Processing and querying large web corpora with the COW14 architecture // Proceedings of Challenges in the Management of Large Corpora (CMLC-3). Mannheim: Institut für Deutsche Sprache, 2015. P.28–34.
- Schäfer R., Bildhauer F. Building Large Corpora from the Web Using a New Efficient Tool Chain // Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul: European Language Resources Association, 2012. P.486–493.
- Scott M. WordSmith Tools (Version 5.0) [Computer Software]. Lexical Analysis Software Ltd., 1996. URL: <http://www.lexically.net/software/index.htm> (accessed on 31.08.2019).
- Sharoff S. Creating General-Purpose Corpora Using Automated Search Engine Queries // WaCky! Working Papers on the Web as Corpus. Bologna: Gedit Edizioni, 2006. P. 63–98.
- Shavrina T., Shapovalova O. To the methodology of corpus construction for machine learning: “Taiga” syntax tree corpus and parser // Труды международной конференции «Корпусная лингвистика — 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. P.78–84.
- Sinclair J. M. Preliminary recommendations on text typology: technical report / EAGLES (Expert Advisory Group on Language Engineering Standards), June 1996.
- Sinclair J. Trust the Text: Language, corpus and discourse. London: Routledge, 2004.
- 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.
- Stubbs M. British traditions in text analysis: From Firth to Sinclair // Text and Technology: In Honour of John Sinclair / eds M. Baker, F. Francis and E. Tognini-Bonelli. Amsterdam: John Benjamins, 1993. P. 1–46.
- Suchomel V., Pomikálek J. Efficient Web Crawling for Large Text Corpora / eds A. Kilgarriff, S. Sharoff // WWW2012. Proceedings of the seventh Web as Corpus Workshop (WAC7). Lyon, 2012. P.39–43.
- Svartvik J. Corpus linguistics comes of age // Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4–8 August 1991 / ed. J. Svartvik. Berlin; New York: Mouton de Gruyter, 1992. P.7–14.
- Svartvik J., Quirk R. A corpus of English Conversation. Lund: Gleerup, 1980.

- Šmerk P. Fast Morphological Analysis of Czech. RASLAN 2009: Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University, 2009. P. 13–16.
- Šmerk P. K počítačové morfologické analýze češtiny. Disertační práce. Brno: Fakulta informatiky MU, 2010.
- Thomas J. Discovering English with Sketch Engine: A Corpus-based approach to language exploration. 2<sup>nd</sup> ed. Brno: Versatile, 2016.
- Tognini-Bonelli E. Corpus Linguistics at Work. Amsterdam: John Benjamins, 2001.
- Tribble Ch. Corpora in ELT: Preliminary results from an internet survey. Paper presented at Teaching and Language Corpora 8. Lisbon, 2008. URL: <http://anafrankenbergsynthasite.com/resources/TaLCLisbon2008Proceedings.pdf> (accessed on 14.06.2019).
- Virtanen T. Corpora and discourse analysis // Corpus Linguistics. An International Handbook. Vol. 2. / eds A. Lüdeling, M. Kytö. Berlin; New York: Walter de Gruyter, 2008. P. 1043–1070.
- Zakharov V. Evaluation and Combining Association Measures for Collocation Extraction // Internet and Modern Society. Proceedings of the International Conference IMS-2017 (St. Petersburg, 21–24 June 2017). ACM International conference proceedings series, ACM Press, 2017. P. 125–134.

# Глоссарий

**Абсолютная 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 переводного текста, выполняемое вручную или автоматически.

**Графематический анализ** — анализ потока символов в текстах на естественном языке, выделение отдельных значимых единиц текста (такенов), возможно, приписывание этим единицам их типов.

**Жанр (регистр)** — одна из классификаций текстов, основанная на таких нелингвистических критериях, как цель создания текста, предполагаемая аудитория, стиль написания и т. д.

**Индексирование корпуса текстов** — составление в автоматическом режиме списков адресов каждого слова текста, индекса.

**Коллокат** — слово (или словоформа), встречающееся в качестве ближнего соседа данного слова (словоформы).

**Коллокация** — регулярное, устойчивое сочетание слов в предложении.

**Коллизия** — 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 (статистику, языковые примеры и др.) по каким-либо лингвистическим параметрам (части речи, грамматической форме, синтаксической функции 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 (динамический)** — постоянно пополняемый и обновляемый корпус текстов, создаваемый в целях мониторинга представляемого корпусом подъязыка или языка в целом.

**Корпус параллельный (переводной)** — двухязычный или многоязычный корпус, который состоит из текстов на одном языке и их переводов на другой (другие) язык (языки).

**Корпус полнотекстовый** — корпус, состоящий из целых текстов, а не из фрагментов.

**Корпус сопоставимый (параллельный)** — набор текстов одной и той же тематической области, написанных независимо друг от друга на двух или нескольких языках.

**Корпус сбалансированный** — презентативный 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 нескольких лет).

**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 образования начальных форм для словоформ.

**Меры (лексической) ассоциации** — статистические методы (формулы)

вычисления степени синтагматической связанности лексических единиц в тексте.

**Метаданные** — данные о данных. В корпусе метаданные представляют собой разнoplановые сведения о текстах: название, автор, пол автора, дата публикации, жанр, тематика и др.

**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 представление ее в виде дерева зависимостей или структуры составляющих.

**Разметка (аннотирование) корпуса** — приписывание текстам корпуса и их компонентам дополнительной информации (метаданных). Метаданные можно поделить на три типа: лингвистические, экстралингвистические, данные о структуре текста.

**Разметка морфологическая** — приписывание компонентам текста морфологической информации (граммем).

**Разметка синтаксическая** — сопоставление предложения на естественном языке, представленного как линейная последовательность словоформ, с его синтаксическим представлением в виде формальной структуры (обычно дерево зависимостей или дерево составляющих).

**Репрезентативность** (представительность, сбалансированность) корпуса текстов — достаточно большой объем корпуса и степень представленности в корпусе всех типов текстов, существующих в описываемом языке (подъязыке).

**Стандарт разметки** — система кодирования метаданных (морфологических, синтаксических и т. п.).

**Теггер (грамматический разметчик)** — программа, выполняющая в автоматическом режиме грамматическую (морфологическую) разметку текстов корпуса.

**Токен** — конкретное слово или другой элемент текста (словоформа, текстоформа, словоупотребление).

**Токенизация** — разделение потока символов в текстах на естественном языке на отдельные значимые единицы (токены).

**Тег (тэг)** — метка, которая присваивается слову или предложению в размеченном корпусе в соответствии с характером разметки.

**Язык (формат) разметки** — система или правила составления разметки машиночитаемого текста. В настоящее время стандартным языком разметки является XML.

**Corpus-based**

Where corpora are used to test preformed hypotheses or exemplify existing linguistic theories. Can mean either:

- (a) any approach to language that uses corpus data and methods.
- (b) an approach to linguistics that uses corpus methods but does not subscribe to corpus-driven principles.

**Corpus-driven**

An inductive process where corpora are investigated from the bottom up and patterns found therein are used to explain linguistic regularities and exceptions of the language variety/genre exemplified by those corpora.

# Список сокращений

АОТ	— «Автоматизированная обработка текста» (название лингвистического интернет-портала)
ИПС	— информационно-поисковая система
КС	— корпусная служба
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)
CQL	— Corpus Query Language
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
<i>ipm</i>	— instances per million
KWIC	— Key Word In Context
LOB	— Lancaster-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-Analysing Computing Tools (название программы)
TEI	— Text Encoding Initiative
USAS	— UCREL Semantic Analysis System
WaC	— Web as Corpus
WaCky	— Web-As-Corpus Kool Yinitiative
WWW	— World Wide Web
XAIRA	— XML Aware Indexing and Retrieval Architecture
XCES	— XML Corpus Encoding Standard
XML	— eXtensible Markup Language

# Предметный указатель

- аннотирование** (*см. также разметка*) 28, 173, 179, 180–186, 228
- база данных** (*база текстов, текстовая база*) 77, 80, 86, 92, 105, 107, 110, 111, 113, 114, 124, 187, 199, 200, 230
- биграмма** 16, 102, 202, 203, 204, 226
- встречаемость** 12, 112, 132, 139, 140, 141, 142, 145, 146, 148, 153, 155, 156, 160, 165, 175, 200
- выравнивание** (*alignment*) 61, 62, 187, 226
- граммема** 34, 39, 76, 112, 140
- графематический анализ** (*см. также токенизация*) 207, 209, 226
- дерево зависимостей** (*грамматическая зависимость, структура зависимостей*) 45, 46, 47, 48, 49, 76, 100, 101
- дискурс:** 10, 14, 65, 106, 158, 167, 168, 169, 175, 176, 180, 194
- жанр** (*см. также регистр*) 13, 19, 22, 23, 24, 25, 33, 34, 54, 55, 56, 57, 58, 70, 73, 74, 77, 97, 98, 108, 116, 119, 128, 132, 135, 173, 180, 192, 226
- запрос** (*язык запросов, типы запросов*) 72, 77, 79, 102, 110, 111, 112, 113, 116, 117, 118, 119, 120, 121, 122, 124, 127, 128, 137, 182, 190
- индекс** (*индексирование*) 59, 79, 80, 110, 114, 226
- интерфейс** 62, 79, 80, 81, 91, 92, 93, 102, 110, 111, 113, 118, 119, 124
- кодировка** (*кодирование*) 26, 28, 29, 38, 40–45, 49, 53, 83, 97, 121, 173, 174, 175
- коллокат** 116, 143, 144, 145, 146, 148, 151, 152, 153, 154, 156, 157, 202, 204, 226
- коллокация** 6, 16, 93, 102, 116, 128, 129, 143, 145, 148, 149, 154, 155, 156, 168, 186, 190, 200, 201, 202, 203, 226
- коллигация** 16, 190, 226
- конкорданс** 17, 19, 79, 82, 91, 93, 112, 116, 117, 118, 121, 122, 123, 128, 135, 136, 138, 139, 143, 148, 160, 186, 226
- контекст** 12, 14, 19, 35, 58, 59, 63, 76, 79, 88, 89, 109, 112, 116, 117, 118, 119, 120, 121, 130, 134, 135, 137, 138, 139, 149, 158, 160, 164, 167, 169, 172, 175, 182, 184, 189, 202
- корпус** (*подкорпус*)
- аннотированный (размеченный): 11, 35, 36, 37, 46, 47, 59, 60, 67, 76, 83, 87, 90, 98, 100, 101, 106, 107, 129, 139, 140, 159, 164, 227
  - веб-корпус: 35, 81, 83, 91, 94, 192, 193
  - выровненный: 227
  - диахронический: 91, 92, 108, 227
  - динамический (мониторный): 57, 59, 227
  - иллюстративный: 57, 58, 186
  - исследовательский: 57, 58
  - исторический: 18, 26, 101, 108
  - морфологический: 57, 204
  - мультимедийный (мультимодальный): 101, 105–107
  - национальный: 10, 22, 26, 41, 55, 56, 70, 85, 86, 87, 88, 90, 93, 96, 98, 99, 100, 103, 105, 113, 127, 134, 190, 205
  - параллельный (переводной): 61–64, 91, 100, 108, 115, 117, 187, 190, 227

- параллельный (сопоставимый): 61, 94, 187, 227
  - прагматический: 185–186
  - репрезентативный (см также репрезентативность): 11, 12, 65, 183, 227
  - сбалансированный (см также сбалансированность): 12, 90, 91, 107, 227
  - семантический: 57
  - синтаксический (treebank): 47, 57, 76, 100
  - синхронический (синхронный): 90, 91, 227
  - специальный (специализированный): 56, 57, 59, 60, 107–109, 187, 203
  - статический: 57, 59
  - устный (звуковой, речевой, фонетический): 31–32, 53, 57, 64–67, 88, 89, 91, 93, 101, 103–107, 135, 159–167, 176–179, 182–185
  - учебный: 67–68, 107, 185–186
- корпуса**
- объем (размер): 12, 21, 27, 33, 34, 58, 59, 65, 68, 73, 74, 80, 81, 82, 85, 86, 87, 89, 91, 92, 93, 94, 95, 96, 97, 99, 101, 102, 105, 107, 112, 115, 128, 139, 141, 142, 157, 159, 181, 190, 193, 201, 228
  - создание: 7, 8, 10, 12, 15, 21, 22, 24, 25, 26, 32, 35, 36, 44, 56, 66, 68, 70–85, 88, 90, 93, 94, 97, 98, 101, 102, 103, 105, 106, 107, 108, 111, 113, 115, 129, 142, 179, 180, 190, 198
  - тип (типология): 10, 22, 23, 56–70, 85, 90, 91, 105
- корпусная лингвистика** 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 26, 27, 28, 30, 32, 56, 58, 73, 93, 95, 117, 118, 129, 131, 133, 135, 146, 168, 180, 190, 193, 205, 206, 227
- корпусный менеджер** 10, 12, 13, 16, 59, 77, 78, 79, 80, 82, 83, 90, 91, 92, 94, 102, 110–131, 143, 159, 186, 201, 227
- корпусный метод** 7, 13, 15, 179, 180, 186, 200, 206
- корреляция (коэффициент корреляции)** 192, 193, 199, 200
- лингвистика** — см. корпусная лингвистика
- лексема:** 12, 47, 50, 71, 76, 93, 97, 112, 121, 129, 130, 134, 140, 189
- лемма (лемматизация)** 34, 35, 40, 41, 42, 43, 44, 48, 49, 63, 73, 75, 76, 78, 97, 112, 115, 117, 119, 120, 128, 190, 191, 192, 195, 198, 199, 201, 227
- менеджер** — см корпусный менеджер
- метод** — см корпусный метод, статистика
- меры ассоциации** 116, 128, 129, 193, 201, 202, 203
- метаданные (метаописание)** 12, 16, 27, 29, 32, 33, 54, 55, 71, 72, 76, 77, 112, 119, 120, 127, 135, 136, 181, 185, 228
- неоднозначность (снятие неоднозначности)** 34, 35, 39, 40, 72, 76, 185
- парсинг:** 45, 76, 109, 228
- парсер (разметчик)** 34, 38, 76, 228
- поиск (поисковая система)** 10, 12, 13, 19, 32, 54, 58, 64, 72, 75, 77, 79, 80, 82, 84, 92, 93, 96, 97, 100, 101, 102, 103, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 124, 128, 133, 149, 159, 179, 180, 183, 186, 187, 190, 193, 195, 204, 205
- разметка (см. также аннотирование)**
- автоматическая: 34, 36, 72, 75, 88, 101, 111
  - анафорическая: 34, 36, 52
  - вертикальная (вертикальный формат): 40, 53
  - дискурсная: 36, 53
  - лингвистическая: 34, 36–53, 56, 69, 81, 83, 84, 182
  - морфологическая: 36–45, 75, 78, 83, 101, 106, 112, 174, 228
  - позиционная: 40–44
  - просодическая: 53

- семантическая: 36, 50–53, 75, 101, 106
- синтаксическая: 28, 34, 36–50, 56, 75, 76, 78, 83, 101, 228
- структурная: 75
- экстралингвистическая: 34, 53–55
- XML-формат: 37–40, 44–45
- разметки языка** 28, 39, 81, 229
- регистр (см. также жанр)** 19, 20, 23, 138, 140, 141, 142, 143, 144, 145, 146, 148, 149, 150, 151, 156, 158, 159, 160, 161, 162, 163, 164, 165, 166, 168, 169, 170, 171, 172, 173, 175, 177, 192, 194, 226
- регулярные выражения** 102, 115, 117, 120, 121–127
- репрезентативность (см. также сбалансированность)** 12, 21–22, 23, 24, 70, 95, 133, 190, 193, 228
- речевой акт** 31, 32, 106, 184, 185
- сбалансированность (см. также репрезентативность)** 21, 23, 24, 82, 89, 190, 192, 228
- связи (парадигматические, синтагматические, синтаксические, семантические, ассоциативные)** 34, 45, 46, 48, 49, 63, 66, 93, 95, 130, 161, 163, 169, 186
- словарь — см. частотный словарь**
- словосочетание** 34, 41, 46, 50, 61, 112, 117, 119, 132, 134, 145, 146, 148, 164, 179, 187, 188, 189, 190, 200, 201, 203
- словоупотребление** 21, 60, 65, 67, 68, 74, 91, 92, 93, 95, 96, 97, 102, 103, 107, 128, 142, 150, 151, 154, 183, 196, 198, 199, 201, 203
- словоформа** 12, 16, 35, 37, 38, 40, 41, 42, 43, 44, 47, 63, 74, 76, 97, 101, 112, 115, 117, 119, 128, 139, 186, 191, 199
- сочетаемость** 129, 130, 131, 134, 137, 138, 143, 156, 157, 201, 203
- стандарт** 26, 27, 28, 29, 30, 32, 36, 43, 44, 45, 121, 228
- статистика (статданные, статистический метод)** 18, 24, 73, 76, 77, 79, 80, 91, 97, 99, 109, 110, 112, 113, 127, 128, 130, 131, 132, 135, 138, 153, 175, 183, 184, 192, 193, 200, 201
- структура** 13, 21, 27, 28, 33, 35, 37, 40, 45, 46, 47, 48, 49, 63, 65, 70, 72, 73, 76, 81, 90, 100, 101, 104, 105, 106, 107, 108, 110, 133, 135, 158, 181
- тер** 30, 31, 32, 33, 34, 38, 39, 41, 43, 44, 45, 48, 49, 50, 51, 52, 53, 55, 60, 68, 76, 88, 97, 120, 128, 228
- теггер (разметчик)** 34, 38, 52, 97, 208, 228
- тезаурус** 16, 129, 130, 149, 187
- токен** 16, 33, 36, 40, 41, 42, 43, 73, 74, 75, 76, 80, 81, 84, 92, 94, 121, 128, 181, 190, 228
- токенизация (см. также графематический анализ)** 71, 74, 75, 78, 83, 228
- формат:** 11, 12, 18, 27, 29, 30, 31, 37, 38, 39, 40, 41, 43, 44, 47, 66, 71, 79, 82, 101, 121, 126, 181, 229
- частота (частотность)**
  - абсолютная: 130, 140, 141, 149, 192, 197, 202, 226
  - относительная: 21, 143, 196, 228
  - нормированная (нормализованная): 140, 141, 142, 144, 150, 228
- частотный словарь (частотный список):** 12, 21, 24, 73, 81, 93, 95, 102, 112, 116, 120, 128, 139, 191, 192, 198, 201
- язык**
  - запросов — см. запрос
  - разметки — см. разметки языка
- CQL:** 119–121, 120, 121, 230
- ipm** 16, 117, 192, 193, 228, 230
- n-грамма** 116, 128
- WaC** 78, 80, 81, 82, 83, 84, 93, 94, 111, 230
- word** 14, 19, 73, 78, 86, 91, 113, 117, 119, 120, 128, 129, 174, 189, 213
- word sketch** 129

**Книги и журналы СПбГУ** можно приобрести:

по издательской цене

в интернет-магазине: **publishing.spbu.ru**

и

в сети магазинов «Дом университетской книги», Санкт-Петербург:

Менделеевская линия, д. 5

6-я линия, д. 15

Университетская наб., д. 11

Набережная Макарова, д. 6

Таврическая ул., д. 21

Петергоф, ул. Ульяновская, д. 3

Петергоф, кампус «Михайловская дача»,  
Санкт-Петербургское шоссе, д. 109.

Справки: +7(812)328-44-22, [publishing.spbu.ru](http://publishing.spbu.ru)

Книги СПбГУ продаются в центральных книжных магазинах РФ,  
интернет-магазинах **amazon.com**, **ozon.ru**, **bookvoed.ru**,  
**biblio-globus.ru**, **books.ru**, **URSS.ru**

В электронном формате: **litres.ru**

---

Учебное издание  
ЗАХАРОВ Виктор Павлович, БОГДАНОВА Светлана Юрьевна  
КОРПУСНАЯ ЛИНГВИСТИКА

Редактор Н. С. Венёва  
Корректор О. С. Каптолъ  
Компьютерная верстка Е. М. Воронковой  
Обложка Е. Р. Куныгина

Подписано в печать 20.01.2020. Формат 60×90  $\frac{1}{16}$ .  
Усл. печ. л. 14,6. Плановый тираж 1000 экз. (1-й завод — 300 экз.). Заказ № .  
Издательство Санкт-Петербургского университета.  
199004, С.-Петербург, В.О., 6-я линия, 11.  
Тел./факс +7(812)328-44-22  
[publishing@spbu.ru](mailto:publishing@spbu.ru)



[publishing.spbu.ru](http://publishing.spbu.ru)

Типография Издательства СПбГУ. 199034, С.-Петербург, Менделеевская линия, д. 5.