ЧАСТЬ 1. ВВЕДЕНИЕ В КОРПУСНУЮ ЛИНГВИСТИКУ

1.1. Основные понятия корпусной лингвистики

Корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с применением компьютерных технологий. Под ***лингвистическим,*** или ***языковым, корпусом текстов*** понимается большой, представленный в машиночитаемом виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач. В настоящее время существует множество определений понятия «корпус». Например, определение, приведенное в учебнике Э. Финегана, гласит: ***корпус*** – репрезентативное собрание текстов, обычно в машиночитаемом формате и включающее информацию о ситуации, в которой текст был произведен, такую как информация о говорящем, авторе, адресате или аудитории [42]. Википедия определяет корпусы как большие и структурированные наборы текстов (теперь обычно в электронном виде), которые используются для статистического анализа и проверки гипотез, проверки случаев встречаемости или обоснования языковых правил по определенным областям [62]. Т. МакЭнери и Э. Вилсон дают следующее определение: корпус – это собрание языковых фрагментов, отобранных в соответствии с четкими языковыми критериями для использования в качестве модели языка [51]. В.В. Рыков определяет корпус текстов как некоторое собрание текстов, в основе которого лежит логический замысел, логическая идея, объединяющая эти тексты и воплощенная в правилах организации текстов в корпус, алгоритме и программе анализа корпуса текстов, сопряженной с этим идеологии и методологии [31].

В приведенных определениях подчеркиваются основные черты современного корпуса текстов – цель («логическая идея»),

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

Целесообразность создания и смысл использования корпусов определяется следующими предпосылками:

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

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

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

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

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

изменения лексического состава языка, проводить анализ лексико-грамматических характеристик в разных жанрах и у разных авторов. Корпусы призваны служить также источником и инструментом многоаспектных лексикографических работ по подготовке разнообразных исторических и современных словарей. Данные корпусов могут быть использованы для построения и уточнения грамматик и в целях обучения языку. Более подробно возможности и примеры использования корпусов в лингвистических исследованиях будут рассмотрены в разделе 3.3.

Сегодня корпусная лингвистика часто понимается как относительно новый подход в лингвистике, который имеет дело с изучением использования языка в «реальной жизни» с помощью компьютеров и электронных корпусов. Корпусная лингвистика имеет, по крайней мере, две черты, дающие ей основание претендовать на положение самостоятельной дисциплины: 1) характер используемого словесного материала; 2) специфика инструментария.

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

Э. Финеган определяет корпусную лингвистику как деятельность, требующуюся для составления и использования корпуса, направленную на исследование естественного употребления

языка [42]. В этом определении подчеркивается созидательная направленность корпусной лингвистики. Двойственный характер корпусной лингвистики (нацеленность как на создание, так и на использование корпусов текстов) обусловливается двойственным характером ее ***объекта*** – корпуса текстов, который, с одной стороны, представляет собой исходный речевой материал для корпусной лингвистики и для других лингвистических дисциплин; с другой стороны, является результатом деятельности корпусной лингвистики.

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

Существует проблема, связанная с терминологией корпусной лингвистики в русском языке, которая пока не установилась в силу следующих причин: ее относительно недавнее происхождение и ее зарождение в США и Великобритании, обусловившее тот факт, что терминология складывалась и продолжает складываться в недрах английского языка. Русские термины, в основном, представляют собой заимствования английских терминов; некоторые из них в других значениях давно существуют в русском языке. Так, русское слово «корпус» стало многозначным задолго до своего появления в качестве термина корпусной лингвистики. Употребление форм этого существительного является проблематичным, поскольку возможны варианты множественного числа «корпусы» и «корпуса». Для значения «массив», которое имеет место в случае языковых корпусов, именительный падеж множественного числа должен быть «ко́рпусы» и, соответственно, прилагательное должно произноситься с ударением на первом слоге – «ко́рпусный» (Большой толковый словарь русского языка, СПб., 1998). В то же время анализ узуса специалистов пока свидетельствует в пользу форм «корпуса́», «корпусно́й», «корпусна́я», которые используются часто, так что можно, видимо, с осторожностью сказать, что в настоящее время этот

вопрос остается открытым. В Приложении 2 приведены некоторые терминологические сочетания и однословные термины, выделенные из корпуса текстов по корпусной лингвистике.

Правила, регламентирующего употребление той или иной формы применительно к корпусной лингвистике, пока нет, хотя, как представляется, победить должен вариант «корпусы», поскольку он отличает терминологическое значение слова от его общеупотребительного значения. В данном учебнике авторы будут использовать именно этот вариант.

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

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

Технологии, которые применяются в корпусной лингвистике, намного старше электронных компьютеров: многие из них коренятся в традиции конца XVIII и XIX веков, 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 [49].

1. Историческая лингвистика: изменения в языке и реконструкция (сравнительно-исторический метод). Одно из главных направлений, повлиявших на современную корпусную лингвистику, пришло из сравнительно-исторического языкознания. Это неудивительно, поскольку лингвисты, занимающиеся историческими исследованиями, всегда использовали тексты или собрания текстов как основные свидетельства. Многие технологии,

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

Многие идеи и технологии, развиваемые с XIX века, были применены и затем развиты корпусной лингвистикой. Составление исторических корпусов по-прежнему представляет большой интерес. Действительно, среди первых корпусов, доступных в электронном виде, были и исторические корпусы.

Появление огромного количества текстов, доступных в электронном формате, сделало возможным относительно быстрый сбор огромного количества данных. Это предоставило возможность лингвистам выиграть за счет статистических методов в лингвистическом анализе, а также разработать и развить новые методы и модели для исследований. Сегодня математически сложные модели языковых изменений могут быть вычислены с помощью данных из электронных корпусов.

2. Написание грамматик, лексикография и обучение языку. Грамматисты XIX века иллюстрировали свои утверждения примерами, взятыми из произведений признанных авторов. Например, Г. Пауль в своей немецкой грамматике использовал произведения немецких «классиков» для иллюстрации каждого своего утверждения – в области фонологии, морфологии и синтаксиса. Сегодня составители грамматик могут также использовать корпусный подход, но теперь корпусы включают не только классику, но и любые другие типы текстов. В частности,

большой интерес проявляется сейчас к грамматике устной речи. В грамматических описаниях языка можно использовать корпусы для получения информации о частотности характеристик использования разных вариантов, регистров и т.д.

Возьмем некоторые ранние примеры из лексикографии. В середине XVIII века, когда С. Джонсон писал 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 были использованы английскими авторами. Во время чтения Джонсон маркировал предложения, контекст которых делал значение слова особенно понятным. Его ассистенты затем выписывали отмеченные предложения на листы бумаги, и Джонсон распределял их для составления и иллюстрации словарных статей в словаре. Проект под руководством сэра Джеймса Муррея (Оксфордский словарь английского языка – OED) потребовал тысячи читателей и полвека для составления.

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

Традиционные школьные грамматики и учебники часто проиллюстрированы искусственно составленными или отредактированными примерами языкового использования. В будущем они мало чем смогут помочь студентам, которые рано или поздно столкнутся с реальными языковыми данными в своих заданиях или в реальном общении. В этом отношении корпусы как источники эмпирических данных играют важную роль в лингводидактике. При обучении языку корпусы обеспечивают

источник для пробуждения у студентов интереса и вовлечение их в самостоятельное изучение аутентичного языкового использования. Важное применение корпусных данных – Computer-Assisted Language Learning (CALL), где основанное на корпусе программное обеспечение используется для поддержки интерактивной учебной деятельности, выполняемой студентами при помощи компьютера.

3. Социолингвистика: языковое многообразие. Вариативная лингвистика началась с составления карт диалектов и сборников диалектных выражений в последней трети XIX века. Ее методы были похожи на методы, использовавшиеся в то время исторической лингвистикой, с одной существенной отличительной чертой: корпусы диалектов систематически составлялись по определенным критериям. Вероятно, это можно рассматривать как предвестник все еще продолжающейся дискуссии о том, что включать в корпус. В настоящее время электронные корпусы часто используются в исследованиях языкового многообразия (например, диалектов, социолектов, регистров). Математические методы (например, мультифакторный анализ) полностью полагаются на доступность таких данных.

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

1.3. История создания лингвистических корпусов

Лингвисты собрали первые корпусы компьютеризированных текстов в 1960-е годы. Первый компьютеризированный корпус –

Брауновский корпус (The Brown Corpus[1]) – включает 500 текстов из американских книг, газет, журналов, впервые опубликованных в США в 1961 году. Каждый текст в Брауновском корпусе имеет длину 2000 слов (имеется в виду словоупотреблений – tokens), и все собрание включает 1 млн. слов (500 текстов по 2000 слов в каждом). Авторы корпуса У. Френсис (W. Francis) и Г. Кучера (H. Kucera) сопроводили его большим количеством материалов первичной статистической обработки: частотным и 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 процедур; с другой стороны, статистика применялась в сочетании с волевыми решениями авторов корпуса, базирующимися на профессиональной интуиции. Для достижения максимальной объективности этого сложного процесса требовалось построение максимально формализованных, прозрачных для проверки и контроля процедур [31].

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

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

интересы лежат в области лингвистики и которые хотят использовать корпус в целях лингвистического описания и анализа.

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 изобретения облегчили процесс создания корпусов, и последние из них содержат уже миллиарды слов (словоупотреблений).

К 1990 году уже было зафиксировано более 600 компьютерных корпусов. По годам составления они были распределены примерно следующим образом [44]:

-1965	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 шли по нарастающей.

Среди современных корпусов английского языка (как британского, так и американского варианта) наиболее известны Британский национальный корпус (British National Corpus – BNC), Международный корпус английского языка (International Corpus of English – ICE), лингвистический Банк английского языка (Bank of English), Корпус современного американского английского (Corpus of Contemporary American English – COCA) и др. В настоящее время корпусы созданы для многих языков мира (см. Приложение 1).

В первой половине 1990-х годов корпусная лингвистика окончательно сформировалась как отдельное направление науки о

языке. «Корпусная лингвистика достигла зрелости» – так Я. Свартвик озаглавил в 1992 году предисловие к материалам первого Нобелевского симпозиума по корпусной лингвистике [60]. Корпусная лингвистика тесно взаимодействует с компьютерной лингвистикой, используя ее достижения и, в свою очередь, обогащая ее.

1.4. Основные характеристики корпусов

1.4.1. Репрезентативность корпусов

Термин «корпус» обычно обозначает собрание текстов конечного фиксированного размера. С течением времени объем и состав корпуса может меняться, однако эти изменения должны либо не менять его структуру, либо менять ее обоснованно. Представительность корпуса, соотношение его отдельных частей (по разным характеристикам) получили название ***репрезентативности***, или ***сбалансировнности***. Объем первых корпусов, как уже говорилось, составлял 1 млн. словоупотреблений (Брауновский корпус, корпус Ланкастер-Осло-Берген, Упсальский корпус русского языка). Такой объем не позволял отражать язык во всем его многообразии. В настоящее время считается, что общеязыковой (национальный) корпус должен включать не менее 100 млн. словоупотреблений. Национальный корпус представляет данный язык на определенном этапе (или этапах) его существования во всем многообразии жанров, стилей, территориальных и социальных вариантов и т. п. (например, НКРЯ, доступный по адресу http://ruscorpora.ru, BNC, ограниченно доступный по адресу http://www.natcorp.ox.ac.uk/ или http://sara.natcorp.ox.ac.uk). Можно сказать, что все современные лингвистические исследования и работы по составлению словарей и грамматик так или иначе ориентированы на использование представительных (репрезентативных) корпусов текстов.

Задача авторов корпуса – собрать как можно большее количество текстов, относящихся к тому подмножеству языка, для

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

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

1. Корпусы первого типа универсальны, они отражают в себе все многообразие речевой деятельности.
2. Корпусы второго типа отражают бытование некоторого лингвистического или культурного феномена в общественной речевой практике, они построены ad hoc (для специальной цели), например, корпус пословиц или корпус политических метафор в газетной речи [31].

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

Речевая действительность чрезвычайно разнообразна, представлена в разных фактурах (устная, письменная, печатная речь и т.д.), и разнообразие зафиксированных в ней лингвистических явлений просто необозримо. В 60-е годы корпусы текстов, относящиеся к *первому типу*, претендовали на то, что они универсальные, то есть отражают статистически корректно всю картину бытования данного языка или некоторый представительный ее фрагмент [51]. Например, Брауновский корпус текстов был создан для отражения печатной речи США 60-х годов с удовлетворительной для того времени степенью репрезентативности. Отобранные тексты, как уже говорилось, должны были представлять 15 жанров (регистров), из которых было сделано от 6 до 80 элементарных выборок:

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

В корпусах *второго типа* критерием репрезентативности будет служить требование максимально объективного представления бытования интересующего его создателей явления. Так, корпус английских пословиц, максимально репрезентативно отражающий их

употребление в речевой практике носителей английского языка определенного времени и географического региона, не будет репрезентативным для изучения, к примеру, английской политической метафоры [31].

В начале XXI века свободно обсуждаются такие корпусы текстов, как корпус газетных заголовков, корпус английских текстов, предназначенных для отладки систем машинного перевода, корпус политических метафор [2]. Очевидно, что здесь критерий отбора текстов для корпуса его создатель задает сам, исходя из целей своей практической или научной деятельности, поскольку в основе корпуса всегда лежит постановка проблемы для проведения научного поиска.

Методология конструирования такого объекта, как корпус, должна зависеть от типа корпуса. Эта проблема является актуальной и недостаточно разработанной. Методология построения корпусов первого типа так или иначе основывается на принципе дедукции – реализации проблемы корректности движения от общего (объективно существующей речевой практики носителей языка) к отражающему это общее частному корпусу текстов. Методология построения корпусов второго типа должна корректно отражать частные, единичные лингвистические феномены в корпусе текстов, специально созданном для их отражения [20]. Теория и практика показывают, что оба эти подхода, тем не менее, часто применяются в комбинированном виде.

1.4.2. Классификация корпусов по различным основаниям

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

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

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

Вообще существует большое число разных типов корпусов, что определяется многообразием исследовательских и прикладных задач, для решения которых они создаются, и различными основаниями для классификации. В зависимости от поставленных целей и классифицирующих признаков, можно выделить различные типы корпусов (*табл. 1*).

Таблица 1

Классификация корпусов

Признак	Типы корпусов
Тип языковых данных	Письменные Устные Смешанные
«Параллельность»	Одноязычные Двуязычные Многоязычные
«Литературность»	Литературные Диалектные Разговорные Терминологические Смешанные
Цель	Многоцелевые Специализированные
Жанр	Литературные Фольклорные Драматургические Публицистические
Доступность	Свободно доступные

	Коммерческие Закрытые
Назначение	Исследовательские Иллюстративные
Динамичность	Динамические (мониторные) Статические
Разметка	Размеченные Неразмеченные
Характер разметки	Морфологические Синтаксические Семантические Просодические и т.д.
Объем текстов	Полнотекстовые «Фрагментнотекстовые»

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

По критерию параллельности корпусы делятся на одноязычные, двуязычные и многоязычные. В одноязычных корпусах противопоставляются диалекты, варианты языка. Например, такие разновидности английского языка, как английский как родной и английский как иностранный оставались за пределами научного интереса до появления новых технологий, позволивших вовлечь в контрастивный анализ существенно большее количество сопоставляемых произведений речи. Двуязычные и многоязычные корпусы объединяют тексты из одной и той же тематической области, независимо написанные на двух или нескольких языках (например, корпус материалов конференций по определенной научной проблеме, проходивших в разных странах и на разных языках). Такие корпусы помогают в работе с терминологией и часто используются переводчиками. Еще один вариант двуязычного или многоязычного

корпуса – множество текстов-оригиналов, написанных на каком-либо исходном языке, и текстов-переводов этих исходных текстов на один или несколько других языков. Такой корпус предоставляет неоценимый материал для проведения сравнительно-сопоставительных исследований, для исследований по теории перевода и для обучения переводу человека и компьютера.

По критерию «литературности» выделяются литературные, диалектные, разговорные, терминологические и смешанные корпусы. Примером разговорного корпуса может быть корпус Один Речевой День (ОРД), разрабатываемый в Санкт-Петербурге [38], примером терминологического корпуса – корпус текстов по корпусной лингвистике, позволяющий разрабатывать терминологический словарь непосредственно на живом текстовом материале [54]. В этом корпусе методология корпусной лингвистики применена к ней самой.

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.

Корпусы текстов могут быть классифицированы *по жанрам* и подразделяться на литературные, фольклорные, драматургические, публицистические и др. Примерами публицистического корпуса могут служить Компьютерный корпус текстов русских газет конца XX-ого века (http://www.philol.msu.ru/~lex/corpus/) и корпус политических метафор [2].

Важным критерием для пользователей корпуса является его *доступность*. Свободно доступные корпусы позволяют в любое время в режиме on-line иметь доступ ко всем текстам корпуса в полном объеме. В ряде случаев свободный доступ может предоставляться к части корпусных данных. В работе с коммерческими корпусами нужно покупать право его использования on-line или копию на компакт-диске. Предварительно можно ознакомиться с аннотацией к корпусу или, возможно, даже

поработать с корпусом в пробном режиме, но, как правило, не со всеми текстами, а только с небольшим по объему подкорпусом. Закрытые корпусы создаются для узко специфических целей и не предназначены для публичного использования.

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 при построении исследовательских корпусов использовать пропорциональное сужение, являющееся наиболее простым способом обеспечения репрезентативности. Как правило, такие корпусы текстов содержат от нескольких десятков миллионов до сотен миллионов словоупотреблений. Иллюстративные корпусы создаются после проведения научного исследования: их цель не столько выявить новые факты, сколько подтвердить и обосновать уже полученные результаты. Они служат для выделения из них лингвистических примеров, подтверждающих те или иные языковые (речевые, текстовые) факты, обнаруженные ранее иными лингвистическими приемами. Типичный пример иллюстративного корпуса представлен в «Путеводителе по дискурсивным словам русского языка» [3], где семантический анализ частиц и выделенные значения сопровождаются значительным текстовым материалом, позволяющим читателю проверить предложенные семантические интерпретации [17; 2].

Критерий *«динамичность»* подразделяет корпусы на динамические и статические. Первоначально корпусы текстов создавались как статические образования, отражающие определенное временно́е состояние языковой системы. Статические корпусы содержат тексты какого-то небольшого временно́го промежутка [17]. Типичными представителями этого вида корпусов являются авторские корпусы – коллекции текстов писателей. Однако значительная часть чисто лингвистических и не только лингвистических задач требует выявления функционирования

языковых феноменов на временно́й шкале – например, изменения значения слов, частоты использования тех или иных синтаксических конструкций и т.д. Для отражения процессуального аспекта проблемной области была разработана новая технология построения и эксплуатации динамического корпуса текстов [2]. Динамические корпусы называют также мониторными или мониторинговыми. Цель мониторных корпусов – «складировать» постоянно растущее количество текстов в памяти компьютера. В течение заранее фиксированного промежутка времени происходит обновление и/или дополнение множества текстов корпуса. Неограниченные (постоянно развивающиеся) мониторные корпусы играют огромную роль в строении словаря, 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. В динамические корпусы текстов, как правило, включают письменные источники большого временно́го периода. Они предназначены для проведения различных диахронических исследований [17].

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

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 относятся некоторые корпусы текстов

определенного автора, а также корпусы коротких текстов, например, корпус мерфизмов (так называемых «законов подлости») [5] или корпус газетных заголовков.

1.4.3.Особые типы корпусов

1.4.3.1. Параллельные корпусы

Параллельные корпусы можно разделить на два основных типа:

1) корпусы, представляющие множество текстов-оригиналов, написанных на каком-либо исходном языке, и текстов-переводов этих исходных текстов на один или несколько других языков;
2) корпусы, объединяющие тексты из одной и той же тематической области, независимо написанные на двух или нескольких языках.

И те, и другие корпусы создаются и используются для сравнительных исследований языков (в области лексикологии, грамматики, стилистики, переводоведения и т.д.), а также в целях разработки эффективных методов перевода, в том числе, машинного.

При подготовке параллельных корпусов текстов первого типа и разработке пакетов программ для их обработки возникает проблема, которая заключается в установлении соответствий между текстом оригинала и его переводами [2]. Для решения этой задачи используется так называемый метод автоматического выравнивания (alignment) текстов. Суть этого метода заключается в параллельной сегментации оригинального текста и его перевода по предложениям, клаузам (грамматическим конструкциям), словосочетаниям и словам. При выравнивании на уровне предложений могут использоваться, как это описано в учебнике А.В. Зубова и И.И. Зубовой [17], шесть возможных соответствий между предложениями обоих текстов.

1) одно исходное предложение переводится одним предложением;
2) два исходных предложения переводятся одним предложением;
3) одно исходное предложение переводится двумя предложениями;

4) два исходных предложения переводятся двумя предложениями, но внутренние границы этих предложений в тексте оригинала и тексте перевода не совпадают;
5) предложение исходного текста не переводится;
6) предложение в тексте перевода не имеет эквивалента в тексте оригинала.

Теоретически обоснованным при решении данной проблемы может быть использование технологий систем машинного перевода с языком-посредником или универсальным языком [2].

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

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

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

В процессе перевода такая система пытается отыскать переводимое предложение или его фрагмент в массиве исходных параллельных текстов. Если оно найдено в исходном массиве текстов-

оригиналов, то система выбирает перевод такого предложения или его части в массиве переведенных текстов [17].

При исследовании параллельных корпусов, в том числе корпусов второго типа, могут успешно применяться инструменты автоматической классификации лексики. Автоматическая классификация лексики является одной из ключевых процедур автоматического понимания текстов [4]. Она осуществляется в рамках формализации 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 второго типа, имеет высокую ценность в определении адекватности перевода и при проведении контрастивных исследований. Применение модулей автоматической классификации лексики повышает эффективность поиска в параллельных корпусах, позволяет извлекать данные для пополнения и корректировки многоязычных словарей, для проверки качества работы систем машинного перевода и их обучения [25; 7].

Система автоматического 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.

Системы переводческой памяти могут быть использованы творчески для большей автоматизации переводческого процесса, не зависящей от конкретных языков. Система машинного перевода основывается на применении синтаксического сходства между более или менее родственными естественными языками. В частности, это касается таких языков, как чешский и словацкий.

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

1.4.3.2. Корпусы устной речи

Прагматика не была так тщательно исследована в компьютерной лингвистике и корпусных исследованиях, как некоторые другие сферы лингвистики, поскольку создание репрезентативного корпуса устной речи было сложной задачей. В конце концов, возникла необходимость создать модели вежливости, смены ролей и других явлений [42].

Составители корпуса не всегда могут представить себе все многообразие лингвистических задач, которые могут быть решены с его помощью. Среди них областью особой важности, основной для понимания языка вообще, является исследование устных текстов. Корпус 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 1960 года под руководством Р. Квирка в Лондонском университетском колледже. Объем корпуса – 1 млн. словоупотреблений. Текстами устной речи были записи радиопередач, заседаний официальных структур, а также неформальных бесед. Машинный вариант корпуса создавался в

Лундском университете (Швеция) и был готов к использованию в 1979 году. Именно корпус устной речи Лондон-Лунд был одним из первых машиночитаемых корпусов. Он состоял из 34 текстов, представляющих тайно записанные разговоры, которые были также опубликованы в книге Дж. Свартвика и Р. Квирка «Корпус английского разговора» (1980) [59]. Эта книга была очень полезна в то время, когда компьютерные корпусы не были широко распространены, и было трудно обращаться со сложной транскрипцией устной речи [44]. Хотя некоторой частью информации пришлось пожертвовать при составлении машиночитаемой версии, и те, кого записали, вряд ли могут считаться среднестатистическими предствителями лиц, говорящих на английском языке, корпус Лондон-Лунд очень помог в изучении речи. Из-за сложностей составления корпусов устной речи этот корпус долго оставался самым важным источником для компьютерного исследования разговорного английского.

Появление корпуса Лондон-Лунд привело к множеству исследований по лексике, грамматике, просодии речи и особенно по структуре и функционированию дискурса. Так, были исследованы использование слов *actually, really, you know, you see, I mean, well,* вопросы и ответы в английском разговоре, использование пассива, просодических моделей английского разговора и т.д. Устный и письменный английский изучались в сопоставительных исследованиях на базе корпусов Лондон-Лунд и Ланкастер-Осло-Берген; в частности, изучались модальность, связи в сложных предложениях, отрицание. В настоящее время большой интерес корпусных лингвистов привлекают способы передачи эмоций в устной речи, выражение удивления и т.д. Примером корпуса, позволяющего проводить подобные исследования, является мультимедийный подкорпус в составе НКРЯ.

Отсутствие баланса в доступности устного и письменного материала в машиночитаемом формате продлится еще очень долго. В силу различных причин, построение корпусов устной речи

продвигается намного медленнее, чем построение корпусов письменной речи. В первую очередь, устную речь нужно как-то зафиксировать – например, с помощью магнитной ленты, цифровой записи или видеокассеты. Затем ее нужно записать буквами, что является утомительной и дорогой работой, качество которой зависит в большой степени от качества записи и степени шума внешней среды в естественных условиях.

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

1. Какой алгоритм использовать для транскрибирования?
2. Учитывать ли индивидуальные особенности произношения?
3. Учитывать ли весь устный текст или его фрагменты?
4. Учитывать ли диалектные варианты произношения слов?
5. Учитывать ли ударения в словах?
6. Учитывать ли просодические признаки произносимых фраз?
7. Отмечать ли слова, которые при прослушивании не распознавались?
8. Отмечать ли в записи для фонетического корпуса паралингвистические явления, сопутствующие речи (паузы, смех, бормотание, кашель, и т.д.)? [17]

В настоящее время общепринято, что для создания машиночитаемых фонетических корпусов используется *транскрипция на основе орфографического представления звуков* речи с дополнительными знаками, передающими (при необходимости) просодические, паралингвистические и другие особенности произношения. Несмотря на трудности создания, в мире уже существует много достаточно представительных фонетических корпусов. Так, как описывается в учебнике А.В. Зубова и И.И. Зубовой, в 70-х годах XX века в США Х. Далем и его коллегами был создан «Корпус устной речи американского варианта английского языка». Он 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 [17].

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

В составе Национального корпуса русского языка (который имеет также название Русский национальный корпус – РНК) в январе 2008 года содержалось всего 3,9% устных текстов. «Устный» компонент корпуса текстов подразделялся на следующие типы: публичная речь (64,3%), непубличная речь (8,1%), речь кино (27,6%) [27].

Таким образом, одной из наиболее важных проблем при составлении национальных корпусов текстов является их недостаточное наполнение устными текстами, особенно относящимися к непубличной речи – телефонным разговорам, неформальным беседам и т.д.