МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«ТЮМЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ИНСТИТУТ СОЦИАЛЬНО-ГУМАНИТАРНЫХ НАУК
Кафедра английской филологии и перевода

РЕКОМЕНДОВАНО К ЗАЩИТЕ В ГЭК
Заведующий кафедрой
д-р. филол. наук, доцент
_____ Н.В. Дрожащих
21.06 2021 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
магистерская диссертация

УЧЕБНЫЙ КОРПУС АКАДЕМИЧЕСКИХ СТАТЕЙ ПО КОРПУСНОЙ ЛИНГВИСТИКЕ: ПРИНЦИПЫ ПОСТРОЕНИЯ, СТРУКТУРА, ЦЕЛЬ И ЗАДАЧИ

45.04.02 Лингвистика
Магистерская программа *«Прикладная лингвистика»*

Выполнила работу
студентка 2 курса
очной формы обучения — Ишимцева Екатерина Дмитриевна

Научный руководитель
канд. филол. наук, доцент — Федюченко Лариса Григорьевна

Рецензент
канд.филол.наук,
доцент кафедры
немецкой филологии — Савина Ольга Юрьевна

Тюмень
2021

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Корпусная лингвистика считается довольно молодой наукой, она появилась лишь во второй половине XX века. Известный отечественный филолог В. А. Плунгян говорит, что «если мы хотим назвать такую область лингвистики, которая по определению является суперсовременной, то первое, что приходит в голову, — это как раз лингвистика корпусов» [URL: https://postnauka.ru/faq/87397].

Корпусная лингвистика занимается созданием корпусов текстов и применением их для решения лингвистических задач, однако «возникнув относительно недавно, она не успела полностью сформироваться, и её статус как независимой науки ещё не установлен» [Майорова, с. 42]. Так, многие учёные рассматривают корпусную лингвистику в качестве подобласти прикладной. Несмотря на это, корпусной лингвистикой к настоящему времени накоплен серьёзный опыт разработки корпусов для различных языков, однако широкое применение последних в лингвистических исследованиях до сих пор отсутствует. Кроме того, в рамках данной науки появилось множество новых терминов, смысл которых не всегда понимают студенты и люди, только начавшие своё знакомство с корпусами. Также среди самих учёных-лингвистов нет закреплённого варианта склонения, к примеру, центрального понятия корпусной лингвистики — корпуса, так как в некоторых учебниках и статьях можно встретить и такой вариант, как «корпусы» или «corpuses» в англоязычных работах. Таким образом, возникает такая **проблема**, как сбор и грамотная систематизация текстов лингвистов, касающихся корпусной лингвистики, а в связи с этим, становится весьма **актуальным** создание учебного корпуса академических статей по корпусной лингвистике, который будет отображать в контексте все основные понятия корпусной лингвистики.

Новизна исследования состоит в том, что впервые в целях отображения функционирования терминологии в области корпусной лингвистики был составлен учебный корпус академических статей по корпусной лингвистике.

Объект исследования — академические статьи по корпусной лингвистике.

Предмет исследования — принципы создания учебного корпуса академических статей.

Таким образом, **целью** исследования является создание и описание учебного корпуса академических статей по корпусной лингвистике.

Для достижения поставленной цели были определены следующие **задачи**:

1) рассмотреть историю корпусной лингвистики,

2) рассмотреть понятие «корпусная лингвистика» и изучить её базовое понятие «корпус»,

3) изучить существующие инструменты для создания корпусов,

4) составить корпус академических статей по корпусной лингвистике,

5) провести терминологическое исследование на базе собранного корпуса.

Материалом исследования послужили англоязычные научные тексты — монографии, сборники статей — по корпусной лингвистике. Всего было использовано 13 источников общим объёмом 999 236 слов, изданных за период с 2001 по 2016 года. Все материалы были взяты из открытых электронных библиотек, таких как Twirpx (twirpx.org) и КиберЛенинка (cyberleninka.ru).

В ходе исследования были применены такие **методы**, как:

1) теоретический анализ и синтез педагогической, методической и научной литературы по теме исследования;

2) методы корпусного анализа лингвистических данных;

3) методы сбора и организации языковых данных в корпус;

4) методы статистической обработки корпусных данных;

и **технологии** корпусной лингвистики для составления учебного корпуса:

1) программы лингвистического аннотирования;

2) программы и сервисы для лингвистической обработки корпусов;

3) корпусно-поисковые системы.

Теоретическую базу исследования составили труды отечественных и зарубежных учёных в области:

- корпусных исследований и преподавания с опорой на корпус: Б. Б. Базарова (2016), Ю. А. Волоснова (2006), И. Ф. Ганиева (2007), Е. В. Грудева (2012), В. П. Захаров (2018; 2011; 2013; 2005; 2015), Е. Калинина (2018), Н. В. Козлова (2013), М. В. Копотев (2014), А. М. Лаврентьев (2004), А. Д. Майорова (2017), В. А. Плунгян (2013), Н. Б. Гвишиани (2008), П. Бейкер (2009; 2006; 2014), Э. Харди (2006), Т. МакЭнери (2001), Р. Факкинетти (2007), Л. Фловердью (2012), С. Кюблер (2015), Дж. Синклэр (1991; 2004) и др.
- прикладной лингвистики: А. Н. Баранов (2001), И. С. Николаев (2016), О. В. Митренина (2016), Э. В. Семенова (2015), Н. Г. Склярова (2016), Е. П. Соснина (2000), Г. Кук (2003), В. Кук (2009), Б. Палтридж (2015) и др.
- компьютерной лингвистики: Е. И. Большакова (2011), К. К. Боярский (2013), А. В. Зубов (2004; 2006), Ю. Н. Марчук (2007), К. Коз (2016) и др.

Научно-исследовательская работа состоит из введения, двух глав, заключения, библиографического списка и двух приложений.

В процессе написания данной работы сформировались высокого уровня способности к самоорганизации и саморазвитию: умение управлять своим временем, умение поддерживать свой уровень физической подготовленности для обеспечения полноценной профессиональной деятельности.

Для успешной подготовки и защиты выпускной квалификационной работы обучающимся использовались средства и методы физической культуры и спорта с целью поддержания должного уровня физической подготовленности,

обеспечивающую высокую умственную и физической работоспособность. В режим рабочего дня включались различные формы организации занятий физической культурой (физкультпаузы, физкультминутки, занятия избранным видом спорта) с целью профилактики утомления, появления хронических заболеваний и нормализации деятельности различных систем организма. В рамках подготовки к защите выпускной квалификационной работы автором созданы и поддерживались безопасные условия жизнедеятельности, учитывающие возможность возникновении чрезвычайных ситуаций.

ГЛАВА 1. ОСНОВНЫЕ ПОНЯТИЯ КОРПУСНОЙ ЛИНГВИСТИКИ

В начале исследования нам необходимо изучить базовый понятийный аппарат корпусной лингвистики, то есть рассмотреть историю корпусной лингвистики, предмет и задачи корпусной лингвистики, а также её центральное понятие — корпус. Таким образом, следуя принципу прямой хронологии, мы представим информацию об истории корпусной лингвистики, расскажем, когда возникла данная наука и как лингвисты составляли корпуса до появления компьютеров, а также представим все этапы развития лингвистики корпусов с момента появления доцифровых корпусов и до наших дней.

Далее, в разделе 1.2. будет проведен сравнительно-сопоставительный анализ определений корпусной лингвистики, которые дают отечественные и зарубежные лингвисты, будут описаны объект и предмет исследования корпусной лингвистики, в каких областях можно применять её методы и какие задачи она решает.

В разделе 1.3 данной главы будет рассмотрено понятие «корпус текстов». Будет описано, что понимается под этим понятием, какие бывают виды корпусов, а также будут представлены основные черты современного корпуса и рассказано, для чего корпуса могут использоваться.

В заключительном разделе будут детально описаны все этапы составления корпусов и программное обеспечение для составления и работы с корпусом.

1.1. ИСТОРИЯ КОРПУСНОЙ ЛИНГВИСТИКИ

Сегодня такие понятия, как «корпус текстов» и «корпусная лингвистика», неразрывно связаны с компьютерами. Мы привыкли осуществлять работу с корпусами в виде компьютерных программ и сайтов, хотя они существовали ещё задолго до изобретения ЭВМ.

Опираясь на результаты анализа литературы по истории развития корпусной лингвистики, мы смогли выделить основные этапы её развития, всего их четыре: этап «доцифровых» корпусов, этап развития лексикографии, этап создания и

использования корпусов в прикладных целях и, наконец, этап появления ЭВМ и, соответственно, развития современных корпусов.

Первым этапом является этап так называемых «доцифровых корпусов», которые составлялись лингвистами ещё задолго до появления компьютеров и задолго до первого употребления термина «корпусная лингвистика» в 1977 году. «Поскольку эти корпуса хранились в неэлектронной форме, а способы обработки данных были неавтоматическими, в истории развития корпусной лингвистики выделяется особый период, называемый доцифровым» [URL: https://dhumanities.ru/?p=667].

В большинстве своём доцифровые корпуса были связаны со священными писаниями разных религий. Самым популярным и исследованным среди них является корпус библейских текстов, состоящий из Ветхого Завета и Нового Завета, которые представляют собой собрания священных для иудаизма и христианства книг, составляющих Священное Писание для данных религий. Именно с этого корпуса впервые появился термин «конкорданс», называемый тогда конкорданцией (или симфонией) и означавший «основанные на Библии списки слов с указанием стихов» [URL: https://dhumanities.ru/?p=667]. Первый конкорданс появился в начале XIII века и назывался «Concordantiae morales sacrae scripturae» («Нравственная конкорданция Священного Писания») [Копотев, с. 12]. Современное же значение данного термина не далеко ушло от первоначального. Так, например, отечественный лингвист Ю. Н. Марчук говорит, что «словарь-конкорданс представляет собой список текстовых употреблений каждого слова, взятых в контексте определенного размера» [Марчук, с. 97]. А доцент кафедры перевода МГУЛ и кандидат филологических наук Ю. А. Волоснова уточняет, что конкорданс это не просто список словоупотреблений или все вхождения определённого слова, а «список словоформ, встречающихся в тексте, расположенных в алфавитном порядке» [Волоснова, с. 48]. А. Б. Кутузов в своем

курсе лекций по корпусной лингвистике подчеркивает главную ценность конкорданса — контекст [Кутузов, с. 39].

Таким образом, зарождение корпусной лингвистики неразрывно связано с религиозными текстами — священными писаниями. Помимо этого, тогда же появилась одна из важнейших отличительных черт, отделяющая корпус от любого другого собрания текстов — конкорданс.

Следующим ярким этапом развития корпусной лингвистики и доцифровых корпусов ученые считают период XVIII-XIX веков, в который активно развивались словари и такая наука, являющаяся разделом языкознания, которая и занималась вопросами и аспектами составления словарей, а также их изучения, как лексикография. Специалист по корпусной лингвистике, адъюнкт-профессор М. В. Копотев пишет, что «многие известные до сих пор словари были созданы авторами на основе многотысячных картотек, по сути – иллюстративных корпусов» [Копотев, с. 12]. В качестве примера самых известных и выдающихся результатов работы с такими картотеками Копотев выделяет словарь американского английского Ноа Вебстера (Webster’s dictionary) и Словарь живого великорусского языка В. И. Даля. Исходя из этого, можно сказать, что в XVIII-XIX веках словари считались корпусами, хотя в наше время между этими понятиями существует большая разница.

Третьим этапом развития корпусной лингвистики считают конец XIX – начало XX века, именно тогда ученые впервые стали создавать корпуса с конкретной целью их применения в лингвистических исследованиях и в прикладных целях. В качестве прикладных, или практических, целей в то время называли более качественное и быстрое обучение языку, как родному, так и иностранному, усовершенствование средств связи, подсчет частотности языковых единиц. Первым таким корпусом, или иными словами частотным словарем, стал Частотный словарь немецкого языка (Häufigkeitswörterbuch der deutschen Sprache), он содержал около 11 миллионов слов и был создан Фридрихом Вильгельмом Кэннингом в

1897 году в Берлине для улучшения стенографической системы немецкого языка. В этот период было составлено большое количество корпусов в виде частотных словарей не только для немецкого языка, но и для других, в том числе и для русского.

Примерно в тот же промежуток времени так называемые лингвисты «нового поколения» начали продвигать идею о том, что необходимо описывать не то, как правильно и как нужно говорить, а то, как в живую говорят носители языка. «Данный принцип описания прежде всего «узуса, а не нормы» и лег в основу современной корпусной лингвистики и сути методологии составления корпусов» [Копотев, с. 13].

Из вышесказанного следует, что данный этап отличается от предыдущих тем, что именно в это время корпуса начали применяться в прикладных целях, а также было положено начало ещё одной отличительной черте современных корпусов — описание «живого» языка.

Последний важный этап развития корпусной лингвистики начался с изобретением первых компьютеров или электронных вычислительных машин (ЭВМ). Однако доцифровые корпуса никуда не исчезли, различные их виды, такие как глиняные таблички, тексты, написанные на бересте, различные бумажные картотеки, продолжали использоваться лингвистами для некоторых исследований и продолжают использоваться в отдельных областях лингвистики и в наше время. До появления и развития такой функции, как аннотирование, новые электронные корпуса отличались от доцифровых лишь форматом хранения и выглядели как просто «аккуратно собранные коллекции текстов» [Копотев, с. 15]. Одними из примеров таких корпусов являются Брауновский корпус 1960-х годов и Упсальский корпус русских текстов 1980х годов, ставший первым русскоязычным корпусом и созданный в Университете Упсалы, Швеция. Аннотацией (аннотированием) же, или разметкой, такие лингвисты, как например И. С.

Николаев и О. В. Митренина называют «приписывание текстам и их компонентам специальных тегов: собственно лингвистических, описывающих лексические, грамматические и прочие характеристики элементов текста, и внешних, экстралингвистических (сведения об авторе и сведения о тексте)» [Николаев, Митренина, с. 145]. В. П. Захаров и И. В. Азарова называют аннотацию «характерной особенностью современных корпусов» [Захаров, Азарова, с. 18]. Зарубежные лингвисты согласны с такими определениями, например бельгийский лингвист Сильвен Грейнджер определяет аннотацию как «добавление пояснительной (особенно лингвистической) информации к существующему корпусу разговорного и/или письменного языка» [Granger, с. 16].

В Россию же термин «корпусная лингвистика» пришёл лишь к 1996 году и впервые был произнесен на лекциях британского лингвиста и одного из создателей знаменитого Международного корпуса английского языка (International Corpus of English) Сидни Гринбаума. По крайней мере, в первый раз данный термин «корпусная лингвистика» появился в русском корпусе и был связан с этим именем: «В декабре народ ломился на лекции по корпусной лингвистике профессора Гринбаума» (журнал «Карьера», № 2, 1999). Трудно сказать, кто из студентов написал эту заметку, но именно она войдет в историю корпусной лингвистики как первый случай письменной фиксации русского термина» [Копотев, с. 12]. Однако корпусная лингвистика в СССР существовала и до этого момента. Так, например, еще в 1985 г. в СССР начались работы по созданию Машинного фонда русского языка под руководством академика А. П. Ершова. «В задачи фонда входило накопление на машинных носителях и в базах данных текстовых, лексикографических и грамматических источников, необходимых для научного изучения русского языка и для осуществления прикладных разработок» [Захаров, 2013, с. 1]. Важным шагом для развития корпусной лингвистики в России стало создание онлайн базы русских текстов Национальный корпус русского языка

(НКРЯ), являющейся целым собранием корпусов. Помимо основного корпуса письменных текстов, НКРЯ включает в себя различные подкорпуса: корпус СМИ 1990-2000-х годов (газетный корпус), корпус устных текстов (корпус живой русской речи), параллельные корпуса письменных текстов и т. д. Общее число словоупотреблений в НКРЯ составляет 364 881 378 слов [URL: http://www.ruscorpora.ru/new/corpora-about.html#task].

В дальнейшем, каждый новый шаг в развитии машинной обработки текстов и другого языкового материала предоставлял всё новые возможности для корпусной лингвистики и создателей корпусов. А с появлением и широким распространением глобальной сети Интернет у каждого человека с выходом в сеть появился не только доступ к корпусам, но и к инструментам для создания собственного корпуса.

Таким образом, несмотря на то, что корпусная лингвистика считается молодой наукой, мы выяснили, что корпуса составлялись ещё за несколько веков до появления первых компьютеров, хоть и выглядели они в большинстве своем как просто собрания текстов или частотные словари, а первые доцифровые корпуса появились от необходимости как-то классифицировать священные писания разных религий. Также мы узнали, что раньше корпуса отражали литературный язык, а идея составлять корпуса на основе «живого» языка стала большим шагом для развития корпусной лингвистики, наряду с изобретением компьютером и переносом корпусов в цифровую форму.

1.2. ПРЕДМЕТ И ЗАДАЧИ КОРПУСНОЙ ЛИНГВИСТИКИ

Некоторые лингвисты считают, что «корпусная лингвистика — это раздел прикладной лингвистики» [Грудева, с. 25], а другие считают, что это — «раздел компьютерной лингвистики» [Захаров, 2005, с. 3]. Однако оба эти утверждения можно считать верными, поскольку «компьютерная лингвистика — направление в

прикладной лингвистике» [Большакова, с. 15]. А прикладная лингвистика, в свою очередь, — «наиболее широкий термин, он включает не только компьютерную и математическую лингвистику, но и лингводидактику и другие науки» [Марчук, с. 36].

Рассмотрим для начала более широкое понятие — прикладная лингвистика, которое в российской и западной лингвистике имеет совершенно разные интерпретации. Так, на западе прикладная лингвистика (applied linguistics, angewandte Linguistik) связывается, прежде всего, с преподаванием иностранных языков, включая методику преподавания, особенности описания грамматики для учебных целей, преподавание языка как родного и иностранного и пр. К примеру, британский лингвист Вивиан Кук считает, что прикладная лингвистика — это «междисциплинарная область исследований и методик, направленных на решение практических задач, связанных с языком и коммуникацией, которые могут быть определены, проанализированы или решены путем применения существующих теорий, методов и результатов работы лингвистов, или путем разработки новых теоретических и методологических основ в лингвистике для решения этих задач» [Cook, 2009, с. 3]. А лингвист Вольфганг Тойберт в своем определении обозначил четкие задачи прикладной лингвистики, а именно «обучение языку, перевод, языковые технологии» [Teubert, с. 110].

Российский же термин берёт своё начало ещё в СССР: он стал широко употребляться в 50-е гг. в связи с разработкой компьютерных технологий и появлением систем автоматической обработки информации. Именно поэтому в русскоязычной литературе вместо термина «прикладная лингвистика» в том же значении часто используются термины «компьютерная лингвистика», «вычислительная лингвистика», «автоматическая лингвистика» [Баранов, с. 8]. В подтверждение этому есть и гораздо более современные определения, например И. С. Николаев и О. В. Митренина описывая актуальные задачи данной научной

области выделяют следующее: «Современные прикладные лингвисты создают компьютерные программы, которые не только помогают людям общаться с компьютером, но и позволяют сделать такое общение более результативным и менее заметным для пользователя» [Николаев, Митренина, с. 9]. Однако нельзя сказать, что отечественные учёные ограничивают термин «прикладная лингвистика» одними лишь компьютерными технологиями, к примеру Н. Г. Склярова даёт более широкое определение: «прикладная лингвистика может быть определена как академическая дисциплина, в которой целенаправленно изучаются и разрабатываются способы оптимизации различных сфер функционирования языковой системы» [Склярова, с. 7].

Изучив вышеприведенные определения, мы можем отметить, что основное различие отечественной и зарубежной трактовок прикладной лингвистики заключается в следующем: западные учёные, определяя сферу деятельности данной научной области, акцентируют внимание в первую очередь на обучение языку, методику преподавания и другие учебные цели. Российские же учёные на первое место ставят задачи прикладной лингвистики, связанные с автоматической обработкой текста, разработкой корпусов и т. д.

Более узким термином считается термин «компьютерная лингвистика». Она, «будучи одним из направлений прикладной лингвистики, изучает лингвистические основы информатики и все аспекты связи языка и мышления, моделирования языка и мышления в компьютерной среде с помощью компьютерных программ» [Соснина, с. 9]. О том, что компьютерная и корпусная лингвистика тесно связаны между собой упоминается, например, в данном определении: «Компьютерная лингвистика как теория представляет собой «раздел … лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с использованием компьютерных технологий» [Ганиева, с. 105]. Однако компьютерная лингвистика занимается не

только корпусами, она «охватывает всё, что связано с использованием компьютеров в языкознании» [Семенова, с. 48].

Западные учёные согласны с данными определениями, например П. Бейкер и Э. Харди определяют данную область как «область лингвистики, которая включает в себя научное изучение языка с точки зрения компьютерного подхода» [Baker, Hardie, с. 41]. Т. МакЭнери в своей монографии «Корпусная лингвистика» говорит о тесной связи компьютерной и корпусной лингвистики и добавляет, что компьютерная лингвистика «изучает способы создания компьютерных систем, каким-либо образом работающих с языком».

И, наконец, рассмотрим более узкое и ключевое для нашего исследования понятие — «корпусная лингвистика». Данная научная область имеет довольно двойственный характер, поскольку нацелена как на создание, так и на использование корпусов текстов. Это обусловливается двойственным характером её 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. Приведённый тезис ясно отражается в определениях как отечественных, так и зарубежных лингвистов.

К примеру, А. М. Лаврентьев считает, что корпусная лингвистика — это не столько направление, а «скорее идеология, согласно которой результаты лингвистического исследования должны опираться прежде всего на анализ текстов (устных или письменных), а не на интуицию исследователя или информанта» [Лаврентьев, с. 121]. А. В. Зубов определяет корпусную лингвистику не как науку, или даже идеологию, а как «цикл исследований, связанных с правилами организации текстов в корпус, разработкой алгоритмов анализа таких текстов в рамках некоторой научной методологии» [Зубов, 2006, с. 23]. В своем определении Н. Г. Склярова подчеркивает созидательную сторону корпусной лингвистики:

«Корпусная лингвистика – раздел языкознания, занимающийся разработкой, созданием и использованием текстовых корпусов» [Склярова, с. 9]. Другую сторону данной научной области, а именно исследования на базе корпусов можно видеть в следующем определении: «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» [Захаров, 2011, с. 9].

Двойственность её характера прослеживается и в определениях зарубежных лингвистов. Так, Э. Финеган определяет корпусную лингвистику как «деятельность, требующуюся для составления и использования корпуса, направленную на исследование естественного употребления языка» [Finegan, с. 58]. С этим определением согласны и П. Бейкер и Э. Харди: «Корпусная лингвистика — отрасль науки, связанная с составлением и анализом корпусов» [Baker, Hardie, с. 50]. Вторую же сторону в своём определении подчеркивают Б. Палтридж и А. Пхакити: «Корпусная лингвистика анализирует корпуса с совокупным объемом до нескольких миллионов слов на устном или письменном языке, что позволяет исследователю устанавливать частоту и вероятность слов или фраз, представляющих интерес, часто также с демографическими характеристиками их пользователей» [Paltridge, Phakiti, с. 384].

Таким образом, исходя из вышеприведённых определений, мы можем сделать вывод, что под корпусной лингвистикой сегодня понимают такую науку, которая занимается как разработкой программного обеспечения для создания корпусов, так и созданием корпусов для различных прикладных целей, а также исследованиями на базе корпусов.

Изучив толкования корпусной лингвистики, перейдем к рассмотрению её задач и методов. Так, первичной задачей корпусной лингвистики считается объективное лингвистическое описание языковой системы, причём к этому

описанию корпусная лингвистика подходит, отталкиваясь от изучения конкретной человеческой коммуникации, от реальных текстов.

В качестве вторичной задачи рассматривается выработка особого способа отражения речевого материала в корпусе текстов. Этот способ, в свою очередь, может использоваться другими лингвистическими дисциплинами. Ещё одна часто выделяемая задача корпусной лингвистики заключается в изучении вероятности лингвистических явлений (в отличие от традиционной лингвистики, которая изучает их (явлений) возможность).

Задачи корпусной лингвистики связаны также и непосредственно с разработкой технологий построения электронных лингвистических ресурсов особого типа – корпусов текстов (Corpora) – и решением задач разного рода на базе этих текстов. В основном, такие коллекции и массивы текстов отражают реальное функционирование того или иного языка, а их перенос в компьютерные среды активизировал их практическое и широкое использование в прикладной лингвистике.

Корпусная лингвистика даёт материал для различного рода исследований языка и его вариантов и определяет основной метод анализа текстов и языка на базе корпусов. Одной из важных особенностей метода анализа на базе корпусов является исследование не только чисто лингвистических явлений (грамматических или лексических функций слов, их связей с другими лексемами), но и таких явлений, как, например, частотности лексем или грамматических конструкций в тех или иных жанрах, диалектах.

Корпусный подход, или метод лингвистического исследования, основанный на корпусах текстов, ориентирован на прикладное изучение языка, его функционирования в реальных средах и текстах, что важно, например, для преподавания языка и для компьютерной лингводидактики [Соснина, с. 75].

Резюмируя всё вышесказанное, мы приходим к выводу, что корпусная лингвистика является подобластью компьютерной и прикладной лингвистики и имеет двойственный характер, занимаясь не только составлением корпусов, но и различными исследованиями на их основе. Данная наука полезна не только учёным-лингвистам, но и школьникам и студентам, поскольку многие крупные корпуса отражают живой язык, что может помочь глубже его изучить в контекстах. С её помощью можно решить множество задач, описанных выше, как, например, изучение вероятности лингвистических явлений или объективное лингвистическое описание языковых систем. Корпусные методы позволяют изучать различные языковые явления, а также функционирование языка в «живой» речи.

1.3. ПОНЯТИЕ «КОРПУС» И ЕГО ВИДЫ

Центральным понятием корпусной лингвистики является понятие «лингвистический корпус», для раскрытия его сущности необходимо определить, что является корпусом в целом.

В широком смысле под корпусом понимается любое собрание текстов. В этой трактовке выделяются размеченные (аннотированные) и неразмеченные корпусы текстов. В качестве неразмеченных корпусов можно рассматривать существующие электронные коллекции текстов: виртуальные библиотеки, архивы электронных версий периодических изданий или новостных лент, которые оказываются достаточными для некоторых исследовательских и учебных целей. Но использование неразмеченных собраний текстов, имеющих инструменты поиска, повышает долю информации, которая может оказаться нерелевантной для исследователя, что значительно затрудняет работу с таким источником. В связи с этим «предметом корпусной лингвистики являются преимущественно размеченные корпусы текстов» [Щипицина, с. 58]. Таким образом, корпус можно кратко охарактеризовать следующим образом:

Корпус = тексты + их разметка.

Понятие корпус текстов, как и большинство лингвистических понятий, не имеет единого общепринятого определения. Авторы впервые созданного в 1963 году корпуса текстов («Брауновский корпус») У. Френсиз и Г. Кучера употребили это понятие в значении «совокупность текстов, считающаяся представительной для данного языка, диалекта или другого подмножества языка, предназначенная для лингвистического анализа» [Зубов, 2006, с. 22]. Этот корпус состоял из 500 отрывков разных текстов печатной прозы США, каждый из которых содержал 2000 словоупотреблений. Они представляли 15 наиболее массовых жанров англоязычной печатной прозы 60-х годов [Зубов, 2006, с. 22].

Сегодня существует несколько взглядов на понятие «корпус» текстов. Так, некоторые лингвисты рассматривают его как совокупность текстов какой-либо определённой категории, жанра или другой объединяющей характеристики. К примеру, К. К. Боярский корпусом называет «некоторое собрание текстов, в основе которого лежит логический замысел, логическая идея, объединяющая эти тексты» [Боярский, с. 26]. Е. И. Большакова также считает, что тексты в корпусе должны быть объединены какой-либо общей характеристикой: «Корпус текстов – это коллекция текстов, собранная по определенному принципу представительности (по жанру, авторской принадлежности и т. п.), в которой все тексты размечены, т. е. снабжены некоторой лингвистической разметкой (аннотациями) – морфологической, акцентной, синтаксической и т. п.» [Большакова, с. 99]. Той же точки зрения придерживается и Н. В. Козлова: «Корпусом считается собрание текстов одного или нескольких языков, связанных между собой определенными параметрами» [Козлова, с. 79].

Другой точки зрения на корпус придерживаются такие отечественные лингвисты как В. П. Захаров и И. В. Азарова. Они считают, что это не просто собрание текстов определённого вида/автора/жанра и т. д., а «уменьшенная модель

языка или, в нашем случае, подъязыка» [Захаров, Азарова, с. 13]. То есть, по их мнению, корпус текстов должен отражать то, как язык функционирует в целом. С данным утверждением согласен и А. В. Зубов: «В сегодняшнем понимании корпус текста — это совокупность текстов, являющаяся достаточной для обеспечения надежных научных выводов о некотором языке, диалекте или ином другом подмножестве языка» [Зубов, 2004, с. 166]. Среди зарубежных лингвистов подобное мнение разделяет, к примеру, С. Кюблер: «Современный лингвистический корпус — это доступное в электронном виде собрание текстов или расшифровок аудиозаписей, которые отобраны для представления определенного языка, языкового варианта или другой языковой области» [Kuebler, Zinsmeister, с. 4]. П. Бейкер добавляет, что такие корпуса могут быть использованы в качестве «эталона, с помощью которого можно «измерить» любое утверждение о языке» [Baker, с. 2].

Другие же учёные считают, что корпус должен служить какой-либо цели. Так, например, утверждает В. П. Захаров: «Под названием лингвистический, или языковой, корпус текстов понимается большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [Захаров, 2005, с. 3]. Подобное определение корпусу текстов даёт и зарубежный лингвист Л. Фловердью: «Ведущие исследователи в области корпусной лингвистики рассматривают корпус, как собрание текстов естественного языка, как письменного, так и устного, составленное для определенной цели» [Flowerdew, с. 3].

Несмотря на разницу в определениях, большинство учёных сходятся во мнении, что современный лингвистический корпус отличает от простого собрания текстов один очень важный элемент — разметка.

Опираясь на вышеприведённые мнения лингвистов, в своей работе мы определяем лингвистический корпус следующим образом: *корпус — это собрание текстов, отобранных по определённым принципам представительности — научной области (корпусная лингвистика) и стилю (научный) — и снабженных лингвистической разметкой, составленное для определённой цели*. В данной работе для нас особенно важны такие характеристики, как область языка и лингвистическая разметка. Под областью мы понимаем стиль, в котором написаны используемые тексты и научную область, в нашем случае речь идёт о научном стиле и корпусной лингвистике соответственно. То есть, нам принципиально важно, чтобы материалы для корпуса были именно о корпусной лингвистике и написаны в научном стиле, дабы он отражал конкретно эту область языка. С помощью лингвистической разметки мы сделаем наш корпус именно лингвистическим, а не просто собранием текстов определённой тематики. Помимо этого, мы составляем наш корпус не просто чтобы отразить функционирование языка в представленной области, но и для прикладных целей, таких как обучение языку и корпусной лингвистике, отражение основных понятий корпусной лингвистики в контексте.

Рассмотрим далее в каких областях возможно применение лингвистических корпусов, какие требования выдвигаются к их составлению и какие типы корпусов существуют.

Говоря об использовании корпусов, лингвисты называют множество различных областей, в которых можно их использовать, например:

1) в лексикографии для создания словарей, определения значения многозначных слов и т. д.;

2) в грамматике для определения частоты морфем, типов словосочетаний и предложений и т. д.;

3) в лингвистике текста для дифференциации типов текста, выявления связей внутри абзаца и между абзацами и т. д.;

4) в автоматическом переводе текстов для поиска контекстов слов, имеющих несколько переводных эквивалентов, поиска переводных эквивалентов в параллельных текстах и т. д.;

5) в учебных целях для выбора цитат, фрагментов произведений, примеров для организации учебных занятий, создания учебных пособий и т. д.

6) в тестировании программ автоматического анализа и синтеза речи и т. д. [Зубов, 2004, с. 166], [Овчинникова, с. 60].

Современные исследователи-лингвисты могут как создавать свои собственные корпусы, а затем проводить необходимые исследования на их базе, так и использовать общедоступные корпусы, созданные другими исследователями и их коллективами.

Таким образом, корпус, представляющий собой размеченное собрание текстов с объёмом слов не менее 100 млн, даёт широкие возможности как для прикладных (работа над принципами автоматической разметки), так и для исследовательских целей.

Рассмотрим далее основные требования к современному корпусу, которые выделяют лингвисты: репрезентативность, полнота, экономичность, структуризация материала, самодостаточность и компьютерная поддержка.

Н. Г. Склярова привела развёрнутые определения данным требованиям: «Под ***репрезентативностью*** понимается способность корпуса текстов отражать все свойства проблемной области, релевантные для данного типа лингвистических исследований, в установленной пропорции, определяемой частотой явления в проблемной области. Если репрезентативность указывает на то, что единицы проблемной области отражаются пропорционально в корпусе данных, то ***полнота*** требует учёта релевантных явлений, даже если это не соответствует идее

пропорционального сужения, поскольку при определенном пороге некоторые явления могут исчезнуть из корпуса. ***Экономичность*** корпуса предполагает, что корпус текстов является не просто строгим подмножеством текстов проблемной области, но, по возможности, существенно отличаться от нее по объему. ***Структуризация материала*** заключается в обеспечении корпуса описью данных, в которых единицы хранения характеризуются по тем параметрам, которые могут оказаться важными для пользователя. ***Самодостаточность*** – свойство не корпуса в целом, а его единиц хранения, на которые могут быть наложены существенные ограничения, т.е. единицей хранения может оказаться не целый текст, а его фрагмент (предложение или группа связанных между собой предложений), который не должен содержать неоднозначности любых типов, например, местоимений, для которых невозможно восстановить антецедент и пр. Что касается ***компьютерной поддержки***, то желательно, чтобы корпус текстов имел комплексное программное обеспечение по обработке данных, обеспечивающих ряд функций» [Склярова, с. 64].

Также до начала создания своего корпуса будет важно остановиться на том, как они классифицируются. Существуют различные виды корпусов, их количество и разнообразие просто огромны, а классифицирующим признаком может выступать «цель создания корпуса, тип языковых данных, «литературность», жанр, динамичность, тип разметки, объем текстов и др.» [Николаев, Митренина, с. 144]. Чаще всего лингвисты выделяют такие виды корпусов, как национальные, параллельные и специальные.

Под ***национальным корпусом*** понимается «собрание текстов в электронной форме, представляющих данный язык на определенном этапе его существования, отображающий данный язык во всем многообразии жанров, стилей, социальных и территориальных диалектов и т. п.» [Волоснова, с. 47]. Такой корпус, являясь как бы лицом языка, представляет его «на определённом этапе (или этапах) его

существования и во всём многообразии жанров, стилей, территориальных и социальных вариантов и т. п.» [Базарова, с. 6]. Самыми известными примеров национальных корпусов являются Национальный корпус русского языка (НКРЯ) и Британский национальный корпус.

Параллельный корпус представляет из себя такой, «в котором тексты являются переводами друг друга» [Кутузов, с. 19], то есть такие корпуса «состоят из текстов и их переводов на другие языки» [Flowerdew, с. 163]. Что касается разметки, то такие корпуса снабжены «метатекстовой, межтекстовой и морфологической разметкой» [Захаров, Азарова, с. 159]. Параллельные корпуса очень удобны в обучении языкам, так как позволяют быстро найти не просто эквивалентное выражение в изучаемом языке, но и контекст.

В нашей же работе нас больше интересуют ***специальные корпуса***. В. П. Захаров определяет их как «сбалансированные корпусы, как правило, небольшие по размеру, подчиненные определённой исследовательской задаче и предназначенные для использования преимущественно в целях, соответствующих замыслу составителя» [Захаров, 2013, с. 6]. «Количество специальных корпусов текстов только для русского языка — это сотни а, возможно, и тысячи наименований: рассказы о сновидениях; русскоязычный эмоциональный корпус; Санкт-Петербургский учебный корпус текстов школьников, изучающих английский язык; Санкт-Петербургский корпус агиографических текстов; Регенсбургский диахронический корпус древнерусских текстов; коллекция древнейших и средневековых славянских и русских текстов «Манускрипт»; рукописные памятники Древней Руси, включая берестяные грамоты, и мн. др.» [Николаев, Митренина, с. 143].

Специальные корпуса создаются для абсолютно любых целей, и каждый может создать корпус для своего конкретного исследования. Несмотря на название «учебный корпус» в формулировке темы нашей работы, мы относим его к

специальному по нескольким пунктам. Во-первых, требования по объёму к специальным корпусам гораздо ниже, чем, например, к национальным, обычно их объём начинается от ~1 млн. токенов, что вполне реально осуществить в рамках нашего исследования. Во-вторых, обычно такие корпуса объединяют тексты по одной теме, в нашем случае — по корпусной лингвистике. В-третьих, специальные корпуса создаются не для отражения языка в целом, а для конкретных целей, включая учебные. Таким образом, вид нашего корпуса — специальный, подвид — учебный.

В данном разделе были представлены различные взгляды учёных на определение корпуса, рассмотрены основные характеристики корпуса, его виды и области применения. Также было определено, как мы понимаем лингвистический корпус именно в нашей работе, какие характеристики нам важны и к какому типу мы относим наш корпус.

1.4. ЭТАПЫ СОСТАВЛЕНИЯ КОРПУСА: ОБЩЕЕ ОПИСАНИЕ

Составление корпуса — довольно трудоёмкая и сложная работа, состоящая из четырех основных этапов, которые мы последовательно рассмотрим в данном разделе:

1) разработка проекта будущего корпуса;

2) сбор, оцифровка и редактирование материалов;

3) разметка;

4) выбор корпусного менеджера.

Разработка проекта будущего корпуса.

Цель данного этапа — определить тематику корпуса, составить план его создания, а также поставить основные вопросы к будущему корпусу, без которых работа существенно замедлится. Итак, в первую очередь важно поставить цель, с которой составляется корпус, и задачи, которые он поможет решить. Исходя из

этого, нужно решить, какие тексты войдут в корпус, какого жанра, типа, времени создания, автора и т. д. Также, опираясь на цели и задачи корпуса, необходимо определить его вид (общеязыковой, параллельный, специальный) и подвид.

Далее, на этапе проектирования необходимо поставить ряд вопросов, которые, в большинстве своем, должны быть решены перед началом составления корпуса: вопрос об объёме корпуса, а также о методике его подсчета.

Учёные выделяют различные единицы корпуса, через которые можно посчитать объём будущего корпуса. Так, «основной единицей корпуса текстов могут быть словоупотребления, основы (корни, леммы) и предложения» [Захаров, Богданова, с. 37]. Другой единицей для измерения объёма корпуса является токен. П. Бейкер определяет токен как «лингвистическую единицу, чаще всего *слово»* [Baker, Hardie, с. 159]. Однако Н. Б. Гвишиани дает данному термину более широкое определение: «токен — это *любая* уникальная строка символов, ограниченная пробелами» [Гвишиани, с. 138]. В. П. Захаров добавляет, что токенами являются не только слова, но и «знаки препинания, числа, другие особые символы также выступают как токены» [Захаров, Азарова, с. 16]. Сам же «автоматический процесс преобразования текста в отдельные токены, например, разделением соединённых слов (he’s), отделением знаков препинания (запятых, точек) от слов и заменой заглавных букв на строчные» называется токенизация [Baker, Hardie, с. 159]. Чаще всего в корпусах показывается информация, как о количестве словоупотреблений, так и о количестве токенов. В нашем корпусе в качестве основной единицы измерения его объема будут использоваться токены.

Что касается необходимого объёма, то в корпусной лингвистике действует правило «чем больше, тем лучше», или «репрезентативнее». Под репрезентативностью понимается «необходимо-достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов и т. п. Имеются разные подходы к определению репрезентативности, можно сказать,

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» [URL: https://myfilology.ru//177/korpusy-i-korpusnaya-lingvistika-osnovnye-ponyatiya/]. То есть, чем больше текстов, чем больше словоупотреблений в корпусе, тем он полнее отражает аспект языка, для которого он был создан. Но объём зависит и от вида корпуса, и если максимум ограничивается лишь технологическими возможностями современных компьютеров, то вот минимальный рекомендованный объём всё же существует. Так, в современной корпусной лингвистике считается, что объём общеязыкового корпуса должен составлять не менее 100 миллионов словоупотреблений. Для специального корпуса оптимальным будет размер около 1 миллиона словоупотреблений. Такой корпус будет считаться хоть и небольшим, но всё же достаточным для конкретных целей и задач.

Помимо этого, на этом же этапе ставятся вопросы, касающиеся материалов корпуса: необходимо решить, что делать с графическими изображениями, формулами и прочей неязыковой информацией, встречающейся в текстах. Чаще всего графические изображения, такие как картинки и графики удаляют из текста и не включают в корпус. Однако, если картинка содержит изображение текста, его можно преобразовать в текстовый формат, поэтому лучше решить, что с ними делать, именно на этапе проектирования, чтобы на следующих этапах такие вопросы не возникали и не тормозили процесс.

Таким образом, в результате первого этапа должны быть четко сформулированы цели и задачи корпуса, определён вид и целевой объём, решено, какие тексты войдут в корпус и какие элементы в них оставлять, а какие убирать.

После того, как проект будущего корпуса составлен и основные вопросы, касающиеся цели корпуса и жанра текстов, решены, можно приступать к

следующему этапу — собственно, сбору, оцифровке и редактированию материалов.

Сбор, оцифровка и редактирование материалов.

Цель данного этапа — отобрать по заданным на предыдущем этапе критериям материалы для корпуса и подготовить их к дальнейшим этапам.

Собрать материалы можно различными способами: купить или найти в библиотеках бумажные источники, найти в интернете электронные тексты, собрать и транскрибировать аудио и видеоматериалы и т. д. Собранные источники должны быть оцифрованы, то есть, переведены в электронный текстовый формат, в случае если это бумажные или видео/аудиоматериалы, а электронные, в свою очередь, должны быть сохранены в формате .txt с кодировкой UTF-8. При отсутствии данной кодировки, большинство программ разметки и корпусных менеджеров не смогут работать с текстами.

Далее оцифрованные источники необходимо редактировать, чтобы они подходили под формат корпуса текстов. Помимо этого В. П. Захаров и И. В. Азарова выделяют еще 4 проблемы, которые необходимо решить на этапе редактирования текстов. «Во-первых, это проблема нормализованного написания слов» [Захаров, Азарова, с. 16]. Сюда входит вопрос о том, нужно ли все слова в тексте приводить к строчной форме написания или оставлять только первые заглавные буквы, чтобы отличать имена собственные; «с другой стороны, наверно, вполне разумно приводить заголовки, написанные заглавными буквами, к строчным, убирать разрядку в заголовках или в тексте, собирая таким образом слова воедино» [Захаров, Азарова, с. 16]. Во-вторых, в одном слове могут использоваться разные алфавиты (call-центр), в тексте могут встречаться слова на иностранных языках. Создатели корпуса должны решить, приводить ли такие единицы к общему виду, то есть переводить, транскрибировать или транслитерировать, либо оставить их как есть, поскольку это языковые явления. В-

третьих, в текстах встречаются «различные буквенно-цифровые последовательности» [Захаров, Азарова, с. 16], например: 4-ого, 6-й. В-четвертых, «обозначения элементов списков при помощи римских или арабских цифр, латинскими или кириллическими буквами, могут восприниматься программами как значимые элементы текста» [Захаров, Азарова, с. 17]. Данный перечень проблем может быть расширен, особенно когда речь идёт о специальных текстах, в которых могут встречаться всевозможные формулы или специфические обозначения.

В результате данного этапа все материалы для корпуса должны быть собраны, оцифрованы, отредактированы в соответствии с принятыми на первом этапе решениями и сохранены в необходимом формате. Таким образом, они будут полностью готовы к следующему этапу — разметка.

Разметка.

Цель данного этапа — «преобразование» собранной коллекции текстов в лингвистический корпус путем добавления лингвистической разметки. Прежде чем мы перейдем непосредственно к способам и видам разметки, необходимо дать этому понятию определение.

Согласно определению бельгийского лингвиста С. Грейнджер, разметка корпуса — «добавление пояснительной (особенно лингвистической) информации к существующему корпусу разговорного и/или письменного языка» [Granger, с. 16]. С данным высказыванием согласен российский лингвист К. К. Боярский, в своём учебном пособии «Введение в компьютерную лингвистику» он пишет: «Разметка — приписывание текстам и их компонентам специальных меток» [Боярский, с. 28]. И. С. Николаев и О. В. Митренина более конкретно определяют приписываемую информацию: «собственно лингвистические, описывающие лексические, грамматические и прочие характеристики элементы текста, и

внешние, экстралингвистические (сведения об авторе и сведения о тексте: автор, название, год и место издания, жанр, тематика и т. п.)» [Николаев, Митренина, с. 145]. Синонимом разметки является аннотация — «общий термин, означающий внесение дополнительной информации к корпусным данным» [Flowerdew, с. 320]. Ю. А. Волоснова отмечает, что «чем богаче и разнообразнее разметка, тем выше научная и учебная ценность корпуса» [Волоснова, с. 47].

В нашей работе разметка трактуется как приписывание текстам и их элементам лингвистической и экстралингвистической информации.

Теперь поговорим подробнее о видах и способах разметки. Первый вид разметки — морфологический, ею занимаются программы, которые называются ***теггеры***. В результате их работы «каждому токену приписываются грамматические характеристики, такие как часть речи, лемма, и набор граммем — род, число, падеж, одушевленность/неодушевленность, переходность и т. д.» [Захаров, Богданова, с. 42]. Такую разметку считают основной, поскольку, во-первых, «морфологический анализ лучше всего автоматизирован, во-вторых, он рассматривается как основа для дальнейших форм анализа — синтаксического и семантического» [Николаев, Митренина, с. 147]. В результате морфологической разметки буду представлены данные в структурированном виде, содержащие теги <text> — текст, <p> — абзац, <s> — предложение, <w> — словоупотребление, <pun> — знак пунктуации. Тег <w> содержит вложенный тег <ana> с атрибутами <lemma> — лемма, <pos> — часть речи, <gram> — набор граммем или морфологических признаков. Для каждой речи существует свой набор признаков, «причем каждое значение признака неявно включает в себя название грамматической категории:

- мн — множественное *число,*
- нс — несовершенный *вид,*

• нп — не*переходный,*

• дст — действительный *залог,*

• прш — прошедшее *время,*

• жр — женский *род,*

• дт — дательный *падеж,*

• пр — предложный *падеж,*

• но — не*одушевленное,*

• кр— *кратное* прилагательное [Захаров, Азарова, с. 52].

Рассмотрим программы для морфологической разметки. Одной из них является программа «TreeTagger», разработанная Хелмутом Шмидом в институте компьютерной лингвистики университета Штутгарта. TreeTagger оперирует деревьями принятия решений и успешно применяется в задачах обработки таких языков, как русский, английский, немецкий, французский, итальянский, испанский, голландский, норвежский, китайский, суахили, латинский, и многих других, включая старофранцузские. Помимо этого данную программу можно адаптировать и под другие языки при наличии словаря и вручную размеченного тренировочного корпуса [URL: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/]. Чарльз Ф. Майер также упоминает такие сервисы, как AMALGAM Tagger – бесплатный теггер, предоставляющий выбор из наиболее популярных систем кодирования или теггеров (tagsets) — Brown, LOB, LLC, SEC, POW, ICE; AUTASYS – автоматический теггер и лемматизатор, использующий тегсеты Lancaster-Oslo-Bergen (LOB) и the International Corpus of English (ICE); Brill Tagger – один из первых теггеров, находящихся в свободном доступе, и, как и TreeTagger, обучаем для работы с любым языком [Meyer, с. 87]. Ещё одним бесплатным теггером является TagAnt. Данную программу разработал профессор факультета естественных наук и

инженерии Университета Васэда, Япония, доктор Лоуренс Энтони. Он занимается разработкой образовательного программного обеспечения, многое из которого предлагает бесплатный доступ, для использования исследователями, учителями и учащимися корпусной лингвистике. TagAnt использует набор тегов от TreeTagger и предлагает пользователям бесплатное использование.

Другим видом разметки является ***парсинг*** — «синтаксическая разметка, приписывание слову или словосочетанию определенных синтаксических признаков» [Базарова, с. 60]. Он выполняется «поверх» данных, полученных в результате морфологической разметки. Парсеры «вычисляют морфологическую и синтаксическую структуры линейных последовательностей слов» [Захаров, Азарова, с. 17]. Они представляют из себя специальные компьютерные программы, облегчающие создателям корпусов работу и анализирующие слова в тексте автоматически, они могут выполнять как морфологический, так и синтаксический анализ [Щипицина, с. 47]. Таким образом, парсеры приписывают каждой синтаксической единице определенные синтаксические признаки, например, их функции в предложении, фиксируют синтаксические связи между словами и словосочетаниями. В отличие от морфологической разметки, синтаксическая менее точна, и после проведения автоматического анализа необходимо редактирование и исправление ошибок специалистом-лингвистом вручную. Поскольку теггинг и парсинг — тесно связанные процессы, многие парсеры имеют и встроенные теггеры, например, the Functional Dependency Grammar of English (the EngFDG parser) и TOSCA Parser, оба имеют похожий функционал [Meyer, с. 91]. В отличие от теггинга, в парсинге нет единой структуры представления результатов анализа, она зависит непосредственно от способа разметки. Существует три основных подхода к синтаксической разметке — грамматика зависимостей, грамматика непосредственных составляющих и

комбинированные теории, например, теория синтаксических групп [Николаев, Митренина, с. 36].

Создателем грамматики зависимостей считается французский лингвист Луи Теньер, и с точки зрения данного метода предложение по структуре похоже на молекулу, поскольку предложение, как и молекула, состоит из элементов и связей между ними. Однако, в отличие от молекулярных связей, связи между элементами в предложении нельзя назвать двусторонними. Они, в большинстве случаев, являются подчинительными — из двух слов одно будет главным, а другое или другие — зависимыми [Николаев, Митренина, с. 36]. Результаты парсинга, выполненного на основе грамматики зависимостей, будут представлены в виде дерева синтагматических отношений между членами предложения, а синтаксически размеченные тексты будут называться treebanks [Склярова, с. 68]. «Деревом» зависимостей такую синтаксическую структуру называют потому, что «главное слово может само стать зависимым к какому-либо другому слову, но только к одному» [Николаев, Митренина, с. 36].

Также кратко опишем подход грамматики непосредственных составляющих. Основным понятием в данном подходе являются собственно непосредственные составляющие (НС), то есть элементы предложения, непосредственно входящие в него, то есть не являющиеся компонентами более крупных его частей. НС-грамматика показывает не только строение предложения, но и деривацию, т. е. способ его порождения. Каждому элементу (словоформе или морфеме) приписывается обозначение его частеречной принадлежности, например, существительное, далее относят его к более крупной группе, существительное — именная группа, и так пока не дойдет до конечной точки — предложения. Такой алгоритм построения синтаксического дерева называется восходящим (Botom-Up): «нижестоящие элементы заменяются на вышестоящие до тех пор, пока для такой

замены не потребуется соединить несколько НС» [Николаев, Митренина, с. 41]. Обратный алгоритм, когда предложение делится на нижестоящие НС до тех пор, пока не дойдет до морфемы, называется нисходящим (Top-Down). Также существуют алгоритмы, предполагающие совмещение двух предыдущих, — комбинированные алгоритмы.

Рассмотрим программы, выполняющие синтаксическую разметку. Одними из самых известных парсеров являются:

1) Лингвистический процессор ЭТАП-3, разработанный ещё в СССР на базе самого успешного компьютерного переводчика того времени. Сегодня это полномасштабный лингвистический процессор, который может анализировать тексты, написанные на русском и английском языке, и самостоятельно строить такие тексты по исходному смысловому заданию [URL: http://iitp.ru/ru/science/works/452.htm]. Он работает в связке с толково-комбинаторным словарем, а потому чаще всего правильно анализирует синтаксически неоднозначные предложения [Николаев, Митренина, с. 50].

2) DictaScope и АОТ — ещё один российский парсер, разработанный компанией Dictum. Поскольку он работает независимо от какого-либо семантического анализатора, чаще возникают проблемы с интерпретацией смысла предложения, чем у ЭТАПа-3.

3) NLTK – инструментарий естественного языка, он представляет собой целый набор библиотек и программ для символьной и статистической обработки естественного языка для английского языка. Он был разработан Стивеном Бердом и Эдвардом Лопером с факультета компьютерных и информационных наук Пенсильванского университета [URL: https://ru.qaz.wiki/wiki/Natural_Language_Toolkit]. Данный парсер использует метод

НС-грамматики, но также позволяет строить версии разбора и в виде деревьев зависимостей [Николаев, Митренина, с. 57].

Помимо вышеупомянутых видов разметки, которые являются основными, существует ещё несколько видов. Один из них — семантическая разметка (semantic annotations). В отличие от прагматики, сосредоточенной на контексте, семантика рассматривает буквальное значение слов, фраз или текста. Существует два уровня семантической разметки: уровень слова (the word level) и уровень смысловых отношений между словами и фразами (the level of semantic relations between words and phrases) [Kuebler, Zinsmeister, с. 83]. Семантическая разметка предусматривает «спецификацию значений слов, разрешение омонимии и синонимии, категоризацию слов (разряды), выделение тематических классов, признаков каузативности, оценочных и деривационных характеристик [Николаев, Митренина, с. 148].

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

1) разряд (например, имя собственное);

2) лексико-семантические характеристики (признаки каузативности, оценки, тематический класс лексемы);

3) деривационные характеристики («отадъективное наречие») [Грудева, с. 61].

Собственно лексико-семантические теги в НКРЯ «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» [Захаров, Богданова, с. 49].

Самой же сложной для автоматической разметки считается анафорическая разметка, которая фиксирует референтные связи между предложениями, в частности, местоименные, связывая их в единый текст. Сложность состоит в том, что большинство систем машинного перевода или автоматической разметки обрабатывает каждое отдельное предложение, отчего и страдает связность выходного текста [Склярова, с. 69]. В результате анафорической разметки личные местоимения помечаются как субъектные (SP) или объектные (OP), указательные как De, притяжательные как PoP и т. д. [Meyer, с. 97].

Также применяется разметка, описывающая ударения и интонацию. Такая разметка называется просодической и используется по большей части в корпусах устной разговорной речи. Часто она сопровождается дискурсной разметкой, которая служит для обозначения пауз, повторов, оговорок и т. д. [Склярова, с. 69].

А. М. Лаврентьев упоминает также филологическую разметку, которая «позволяет включать в корпус варианты текста, авторскую и редакторскую правку, выделять иностранные слова, цитаты, прямую речь персонажей литературного произведения, разного рода стилистические фигуры» [Лаврентьев, с. 124].

Все вышеперечисленные способы разметки относятся к лингвистической разметке, однако есть и разметка экстралингвистическая. Она включает в себя такие метаданные, как библиографические характеристики, типологические характеристики, тематические характеристики, социологические характеристики,

«формальную» структурную разметку (текст, раздел, глава, часть, абзац, предложение, а также даты обработки текстов, источники электронных версий, исполнителей. Помимо этого такая разметка может содержать и информацию о жанровой принадлежности текста, о сфере функционирования (бытовая, официально-деловая). Метаразметка необходима для обнаружения связей между языком и условиями его существования, а также для изучения отдельных подмножеств языка [Николаев, Митренина, с. 148], [Захаров, Богданова, с. 51]. В корпусе она выполняет следующие функции:

- служит для формирования архитектуры корпуса;
- позволяет контролировать процесс информационного наполнения корпуса, оценивать его представительность и сбалансированность;
- обеспечивает возможность поиска и отбора текстов пользователем для составления подкорпусов с заданными свойствами [Савчук, с. 62].

Существует подробная схема для метатекстовой разметки, которую предлагает НКРЯ:

I. Информация об авторе

1. Имя автора:
   - конкретный автор, его имя и фамилия;
   - обобщенный автор, если текст создан от лица организации, органа власти, печатного органа и пр.;
   - коллективный автор, такой тип авторства имеют коллективные монографии, совместные публикации.
2. Пол автора.
3. Возраст автора:

указывается только год, если известен, не указывается для коллективного автора.

II. Информация о тексте

1. Название текста.

2. Дата создания текста.

3. Размер текста:

указывается общее количество словоупотреблений в тексте.

4. Сфера функционирования текста:

- учебно-научная;
- производственно-техническая;
- официально-деловая;
- публицистика;
- реклама;
- церковно-богословская;
- художественная;
- бытовая.

5. Тема текста или предметная область.

6. Тип текста:

например заметка, репортаж и интервью в публицистике, научная статья, реферат и учебник в учебно-научной сфере и т.д.

7. Жанр текста:

используется для характеристики художественных текстов.

8. Стиль текста.

III. Информация об аудитории

1. Возраст аудитории:

если не имеет значения, ставится пометка «н-возраст».

2. Уровень образования аудитории:

При метаразметке текста используются четыре значения этого параметра: 1) «высокий» уровень образования, если текст предназначен для читателя с высоким уровнем общего

образования и общим знанием о предмете; 2) «профессиональный», если текст предназначен для специалистов в данной области; 3) «низкий», если текст предназначен для читателя с низким уровнем общего образования и отсутствием специальных знаний о предмете; 4) в остальных случаях используется помета «н-уровень».

IV. Библиографическое описание текста

1. Источник текста:

используются выходные данные книги.

2. Тип носителя:

в какой форме существовал текст до включения его в корпус. [Савчук, с. 66].

Таким образом, в результате данного этапа собранные ранее тексты приобретают один или несколько видов лингвистической разметки и экстралингвистическую разметку. Несмотря на такое множество видов, В. П. Захаров считает, что многие из существующих сегодня корпусов используют либо морфологическую, либо синтаксическую разметку, либо же комбинируют их. Исходя из этого, основными типами разметки можно назвать морфологическую (теггинг) и синтаксическую (парсинг) [Склярова, с. 69].

Теперь остановимся на последнем этапе — выборе корпусного менеджера.

Выбор корпусного менеджера.

Цель данного этапа – выбрать подходящее программное обеспечение, которое бы позволило полноценно использовать корпус: строить конкордансы, давать статистическую информацию по отдельным элементам корпуса и т. д.

Для начала рассмотрим само понятие такого ПО — корпусного менеджера. Ю. А. Волоснова определяет корпус-менеджер как «лингвистическую поисковую систему, обеспечивающую удобный интерфейс для пользователей, другими

словами, корпус-менеджеры — поисковые системы на базе определенного корпуса» [Волоснова, с. 48]. В результате работы корпусного менеджера пользователь получает необходимую статистическую информацию в готовом виде [Базарова, с. 7]. Эту статистическую информацию корпусные менеджеры обычно предоставляют в виде конкорданса, поэтому их ещё называют конкордансерами. В широком смысле слова, «конкорданс выглядит как список всех вхождений определенного искомого термина или запроса в корпусе, представленных в контексте» [Baker, с. 71]. Именно то, что конкорданс дает контекст слова, и представляет особую ценность [Кутузов, с. 39]. Часто конкорданс помимо контекста даёт и ссылки на источники для найденных вхождений [Николаев, Митренина, с. 141]. Также полученные статистические данные могут включать в себя «частотные характеристики отдельных языковых единиц, или граммем, или могут характеризовать совместную встречаемость нескольких лексических единиц» [Склярова, с. 71].

Лингвисты выдвигают к современному корпусному менеджеру ряд требований. Он должен:

— строить как KWIC (Key Word In Context), так и полные конкордансные списки;

— искать контексты не только по отдельным словам, но и словосочетаниям;

— осуществлять поиск по шаблонам (сложные запросы);

— сортировать списки по нескольким критериям, выбираемым пользователем;

— давать возможность отображать найденные словоформы в неограниченном контексте;

— давать статистическую информацию по отдельным элементам корпуса;

— отображать леммы, морфологические характеристики словоформ и метаданные (библиографические, типологические), что зависит от степени размеченности корпуса;

— сохранять и распечатывать результаты;

— работать как с отдельными текстами, так и с корпусами неограниченных размеров;

— быстро обрабатывать запросы и выдавать результаты;

— поддерживать различные форматы текстовых данных (txt, doc, rtf, html, xml и др.);

— быть лёгким (интуитивно понятным) в использовании, как для опытного, так и для начинающего пользователя [Захаров, Азарова, с. 19], [Николаев, Митренина, с. 150], [Захаров, Богданова, с. 55].

Такой набор «умений» считается базовым, однако некоторые корпусные менеджеры позволяют выполнять и более сложные задачи, например «выявлять коллокации (устойчивые сочетания), ключевые слова и словосочетания, строить лексико-семантические группы, частотные списки и т. д.» [Захаров, Азарова, с. 20].

Лингвисты говорят о четырёх поколениях корпусных программных средств и большинство современных корпусных менеджеров относятся к третьему поколению. Они имеют функционал, позволяющий выполнять вышеописанные операции, многоязычную поддержку и интуитивно понятный интерфейс. Примеры таких программ — WordSmith Tools, MonoConc Pro, ParaConc, AntConc [Николаев, Митренина, с. 151].

Рассмотрим подробнее один из корпус-менеджеров четвёртого поколения — Sketch Engine, который мы будем использовать при создании своего корпуса. Он был разработан в 2003 году компанией Lexical Computing Limited, основанной лексикографом Адамом Килгаррифом. По его словам, целью данного корпуса

является охват как можно большего количества языков [Kilgarriff, с. 17]. В его состав входит более 400 лингвистических корпусов по 70 языкам мира, но что нас заинтересовало больше всего — Sketch Engine предоставляет пользователям возможность составить свой собственный корпус, причем как на материалах текстов из интернет-ресурсов, так и на своих материалах.

Создание корпуса на базе корпусно-поисковой системы Sketch Engine происходит следующим образом. Система предлагает два варианта составления корпуса: на материалах текстов из сети Интернет и на материалах пользователей. В первом случае необходимо ввести от 3 до 20 ключевых слов или фраз по теме будущего корпуса, либо ввести конкретные ссылки на необходимые тексты, либо же ввести адреса интересующих ~~вас~~ сайтов, в таком случае в корпус будут выгружены все тексты с данного сайта. Далее система сама соберет тексты, разметит их и предоставит готовый корпус. Во втором случае пользователю необходимо самому собрать интересующие его тексты, отредактировать их при необходимости и загрузить в систему. Следующим шагом Sketch Engine автоматически разметит тексты, используя морфологическую разметку. В этом и заключается его главная особенность и удобство — он функционирует как теггер и корпусный менеджер одновременно. Что касается списка тегов, Sketch Engine тоже использует набор тегов от TreeTagger, но со своими модификациями, такими как:

1) токен «to» может помечаться тегом IN (предлог, подчинительный союз), если это предлог или тегом TO, только если это маркер инфинитива.

2) неопределённые артикли «a/an» оба лемматизируются как «a».

Увидеть фрагмент данного списка тегов можно в Таблице 1, с полным списком тегов можно ознакомиться в Приложении 1.

Таблица 1.

Список тегов TreeTagger с модификациями от Sketch Engine [Sketch Engine, URL]

POS Tag	Description	Example
CC	Coordinating conjunction	and
CD	Cardinal number	1, one
CDZ	Possessive pronoun	one’s
DT	Determiner	the
EX	Existential there	There is
FW	Foreign word	d’hoevre
IN	Preposition, subordinating conjunction	In, of, like
IN/that	That as subordinator	that
JJ	adjective	green
JJR	Adjective, comparative	greener

Далее размеченные тексты система собирает в полноценный корпус и предоставляет полностью весь свой функционал, как корпус менеджер.

Sketch Engine использует стандартные инструменты работы с корпусом, такие как Word Sketch, Word Sketch Difference, Thesaurus, Wordlist, N-grams, Key Words, Text Type Analysis, OneClick Dictionary и, конечно, Concordance. Система предлагает строить конкордансы по лемме, фразе, слову, символу, а также Sketch Engine использует язык запросов CQL (The Corpus Query Language). CQL – это код, используемый для установки критериев поиска в корпусе сложных фраз, которые не могут выполняться с использованием стандартных элементов управления пользовательского интерфейса. Когда мы вводим в стандартную строку поиска лемму или фразу, Sketch Engine автоматически переводит её в CQL [Thomas, с. 91]. Выглядит это следующим образом, пример взят из инструкции Sketch Engine по составлению CQL-запросов: [lemma="drive"] [lc="my"] [lc="own"]? [lemma="car"]. С помощью данного запроса мы найдем примеры выражений «drive my car» или «drive my own car». Таким образом, корпусно-поисковая система Sketch Engine предоставляет не только широкие возможности для работы и анализа на базе уже существующих корпусов, а также делает

возможным для любого пользователя создать свой корпус под интересующие цели и задачи, предоставляя весь инструментарий: разметку и корпус-менеджер.

В результате четвёртого этапа получается корпус, полностью готовый к использованию, на базе которого уже можно проводить различные исследования, выделять ключевые термины, строить коркондансные списки и т. д.

Резюмируя всё вышесказанное, в данном разделе мы подробно описали процесс создания корпуса, начиная от постановки цели и заканчивая готовым корпусом. Были подробно рассмотрены все основные этапы составления корпуса текстов, а именно проектирование будущего корпуса, сбор, оцифровка и редактирование материалов, разметка и выбор корпусного менеджера. Также было представлено различное программное обеспечение для составления корпуса на разных этапах.

ВЫВОДЫ ПО ГЛАВЕ 1

В Главе 1 мы рассмотрели основные понятия корпусной лингвистики и её историю. Мы узнали, что, несмотря на то, что корпусная лингвистика как наука появилась лишь во второй половине прошлого века, учёные создавали так называемые «доцифровые» корпуса ещё задолго до появления компьютеров. В основном эти корпуса являлись простыми собраниями текстов определённой тематики, чаще всего религиозной, или различные словари. Именно из религиозных текстов тех времен появилось понятие, которое сегодня является базовым для лингвистического корпуса текстов — конкорданс.

Помимо этого мы выяснили, что раньше корпуса отражали только литературный язык, а идея составлять корпуса на основе «живого» языка, которые бы отражали узус, а не норму, стала прорывной для корпусной лингвистики, наряду с переносом корпусов в цифровой формат с изобретением ЭВМ.

Далее мы изучили три смежных понятия — «прикладная лингвистика», «компьютерная лингвистика» и «корпусная лингвистика». Проанализировав

определения учёных, мы установили, что прикладная лингвистика является наиболее широким понятием из трёх представленных, а корпусная лингвистика — наиболее узким. Сравнив определения прикладной лингвистики отечественных и зарубежных лингвистов, мы выявили интересное различие: зарубежные учёные первостепенными задачами прикладной лингвистики ставят обучение языку и всё, что с ним связано. Отечественные же учёные на первое место выводят технические задачи, такие как разработка компьютерных программ, связанных с языком, составлением корпусов, исследования на базе корпусов и т. д.

Взгляды учёных на предмет и задачи корпусной лингвистики тоже немного различаются, исходя из чего, мы сделали вывод о двойственном характере данной научной области. Так, корпусная лингвистика занимается не только всевозможными исследованиями на базе существующих корпусов текстов, но и непосредственно разработкой корпусов, разработкой программного обеспечения для этого, изучением принципов их построения.

Также мы рассмотрели центральное понятие корпусной лингвистики — «корпус» текстов. Мнения учёных на счёт него также разделились. Одни лингвисты акцентируют внимание на том, что тексты в корпусе должны быть связаны единой характеристикой, будь то жанр, автор или языковая область. Другие называют корпус уменьшенной моделью языка, утверждая, что он должен отражать функционирование языка или какой-либо его подобласти или раздела. Третьи вовсе главной чертой лингвистического корпуса называют конкретные прикладные задачи, которые можно решить с его помощью. Однако все сходятся во мнении, что наиважнейшей чертой лингвистического корпуса является разметка. Опираясь на определения лингвистов, мы вывели собственное определение, на которое будем опираться при составлении своего корпуса. Наше понимание корпуса ближе к первой группе ученых.

Помимо этого, мы определили, что корпуса текстов используются в различных областях и бывают разных видов, поэтому их можно классифицировать по множеству различных критериев: по цели создания, по типу языковых данных, жанру, типу разметки, объёму текстов и т. д. Опираясь на представленную классификацию, мы определили вид нашего корпуса по следующим критериям: по цели создания — учебный, по типу языковых данных — специальный.

В последнем разделе были рассмотрены и подробно описаны этапы составления корпусов:

1) Разработка проекта. Мы выяснили, что данный этап необходим для того, чтобы определить тематику корпуса, его цели, предполагаемый объём, а также решить, какого рода материалы войдут в него.

2) Сбор, оцифровка и редактирование материалов. Было определено, что данный этап необходим, чтобы отобрать материалы по критериям, определенным на первом этапе, отредактировать их и конвертировать в нужный формат, чтобы они были готовы для работы на следующих этапах.

3) Разметка. Данный этап нужен, чтобы сделать из обычного собрания текстов лингвистический корпус. Мы выяснили, что существует множество видов разметки: морфологическая, которая считается основной, синтаксическая, семантическая, анафорическая, просодическая, филологическая, экстралингвистическая. Также мы подробно рассмотрели для чего нужен каждый из этих видов разметки и какое ПО существует для её выполнения.

4) Выбор корпусного менеджера. Данный этап является заключительным и необходим для того, чтобы использовать собранные и размеченные материалы, как полноценный корпус. Мы узнали, что ученые выдвигают ряд требований к современным корпусным менеджерам, например, умение строить KWIC и конкордансные списки, осуществлять сложный поиск по шаблонам, таким как CQL, давать статистическую информацию по отдельным элементам и т. д. Также

мы выяснили, что лингвисты выделяют 4 поколения корпусных менеджеров, большинство современных из которых относятся к третьему поколению, а корпус-менеджеры четвёртого поколения позволяют работать с большим объёмом данных, поскольку хранит их на серверах.

Также мы подробно остановились на одном из таких корпус-менджеров — Sketch Engine, на базе которого составили наш корпус. Были описаны принципы его работы, используемая разметка, возможности составления на его основе пользовательских корпусов.

ГЛАВА 2. СОСТАВЛЕНИЕ И ОПИСАНИЕ УЧЕБНОГО КОРПУСА ПО КОРПУСНОЙ ЛИНГВИСТИКЕ

В Главе 2 мы подробно остановимся на описании процесса создания учебного корпуса по корпусной лингвистике. В разделе 2.1. будет поэтапно описано составление учебного корпуса по корпусной лингвистике, от постановки цели и отбора материалов, до разметки и готового корпуса. В разделе 2.2. мы проведем терминологическое исследование на его базе.

2.1. УЧЕБНЫЙ КОРПУС АКАДЕМИЧЕСКИХ СТАТЕЙ ПО КОРПУСНОЙ ЛИНГВИСТИКЕ: ОПИСАНИЕ ЭТАПОВ СОСТАВЛЕНИЯ

В данном разделе мы опишем процесс создания учебного корпуса академических статей по корпусной лингвистике в соответствие с этапами, описанными в предыдущем разделе.

Итак, ***первый этап — проектирование корпуса***. В данном исследовании было решено составить учебный корпус академических текстов по корпусной лингвистике, целью которого является отражение основных понятий данной научной области в контексте трудов ученых-лингвистов. Основываясь на цели, было решено включить в корпус такие материалы, как англоязычные учебные пособия, статьи, монографии, посвященные корпусной лингвистике. Поскольку в наш корпус войдут материалы конкретной предметной области, а также опираясь на формулировку цели, вид корпуса мы определяем как специальный. На этом же этапе был решен вопрос о неязыковой информации. Было решено не включать в корпус такие элементы, как содержание, номера страниц, графики и рисунки, и, соответственно, их названия, примеры в виде отрывков на других языках помимо английского. Корпус было решено сделать англоязычным, не параллельным, объёмом не менее 1 миллиона токенов.

На втором этапе мы занимались сбором материалов, их оцифровкой и редактированием. Материалами послужили англоязычные академические тексты — монографии, сборники статей и учебные пособия — связанные непосредственно с корпусами и корпусной лингвистикой, её проблемами и методами, а именно:

1) Paul Baker, Using Corpora in Discourse Analysis;
2) Roberta Faccinetti, Corpus Linguistics 25 years on;
3) Eileen Fitzpatrick, Corpus Linguistics Beyond the Word;
4) Lynne Flowerdew, Corpora and Language Education;

5) Gvishiani N. B., English on Computer, a Tutorial in Corpus Linguistics;

6) Sandra Kübler, Corpus Linguistics And Linguistically Annotated Corpora;

7) Anke Lüdeling, Corpus Linguistics An International Handbook;

8) Tony McEnery, Corpus Linguistics: Method, Theory and Practice;

9) Charles F. Meyer, English Corpus Linguistics: An Introduction;

10) Antoinette Renouf, Corpus Linguistics: Refinements and Reassessments;

11) Thomas James, Sketch Engine: a Toolbox for Linguistic Discovery;

12) Elena Tognini-Bonelli, Corpus Linguistics At Work;

13) Wolfgang Teubert, Text Corpora and Multilingual Lexicography.

Все материалы были взяты в открытом доступе в сети Интернет, на таких ресурсах, как twirpx.org и cyberleninka.ru, представляющих из себя бесплатные электронные библиотеки. Поиск материалов осуществлялся по тегам «corpus linguistics», «corpora», «linguistic corpora», из них были отобраны тексты, посвященные именно описанию корпусной лингвистики, её истории, методов, проблем и задач, а также описанию корпусов текстов, принципов их построения и т. д. Следом за этим, для отобранных материалов была выполнена метатекстовая разметка (метаразметка) или, другими словами, были составлены «паспорта» текстов. При выполнении метаразметки текстов, вошедших в наш корпус, мы опирались на схему от НКРЯ, описанную в предыдущем разделе. Ввиду размеров таблицы, было решено вставить в текст её фрагмент, с которым можно ознакомиться ниже в Таблице 2. Полный вариант таблицы можно найти в Приложении 2.

Таблица 2.

Метаразметка текстов, вошедших в корпус

Информация об авторе	**Информация о тексте**	**Информация об аудитории**	**Библиографическое описание текста**
• Paul Baker	• Using Corpora	• Н-возраст	• LexisNexis, a

Информация об авторе	Информация о тексте	Информация об аудитории	Библиографическое описание текста
• мужской • 1972г.	in Discourse Analysis • 2006 • 118 908 словоупотреблений • учебно-научная сфера • Корпусная лингвистика • монография • научный стиль	• профессиональный уровень	division of Reed Elsevier Inc. • Электронная версия
• Roberta Faccinetti • женский • 1967г.	• Corpus Linguistics 25 Years on • 2007 • 112 771 • учебно-научная сфера • корпусная лингвистика • монография • научный стиль	• Н-возраст • профессиональный уровень	• Editions Rodopi B.V., Amsterdam – New York • Электронная версия
• Eileen Fitzpatrick • женский	• Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse • 2007 • 85 096 словоупотреблений (отрывок) • учебно-научная сфера • корпусная лингвистика • монография • научный стиль	• Н-возраст • профессиональный уровень	• Editions Rodopi B.V., Amsterdam – New York • Электронная версия
• Lynne	• Corpora and	• Н-возраст	• Palgrave

Информация об авторе	Информация о тексте	Информация об аудитории	Библиографическое описание текста
Flowerdew • женский	Language Education • 2012 • 94 791 • учебно-научная сфера • корпусная лингвистика • монография • научный стиль	• профессиональный уровень	Macmillan • электронная версия

Как видно из таблицы, большинство характеристик, касающихся информации о тексте и аудитории, одинаковы для всех источников. По половому признаку получилось следующее: 5 работ, написанных мужчинами-лингвистами, 6 работ, написанных женщинами-лингвистами, и 2 работы с коллективным автором разных полов. По дате публикации: 7 текстов являются современными, 3 новыми и 1 новейшим. Оставшиеся 2 материала являются наиболее ранними из представленных, они были изданы в 2001 и 2004 годах.

Выполнив метаразметку и проанализировав отобранные источники, мы выполнили следующий пункт на данном этапе — оцифровку и редактирование текстов. Несмотря на то, что все материалы были получены уже в электронных версиях (в формате .pdf), они ещё не были готовы к тому, чтобы из них собрать корпус, поскольку, как было сказано в Главе 1, лингвистический корпус текстов — это коллекция текстов, которые имеют лингвистическую разметку. Для того чтобы программы лингвистической разметки (теггеры) смогли открыть и обработать наши тексты, они должны быть сохранены в формате .txt с кодировкой UTF-8, а для этого они должны сначала быть сохранены в формате .docx. Существует два способа сделать это — простой и очень быстрый, с помощью онлайн-сервисов для конвертирования документов из одного формата в другой, и потруднее и очень затратный по времени, то есть вручную. Поскольку в любом случае перед

разметкой тексты необходимо подготовить, т. е. отредактировать, был выбран второй вариант. Так как автоматически удалить номера страниц при копировании документа целиком не представлялось возможным, было решено переносить каждую книгу из .pdf в документ Word по одной странице, что заняло очень много времени, но позволило выполнить вместе с тем сразу и редактирование. Как было решено на предыдущем этапе, мы удалили из текстов такую неязыковую информацию, как содержание, номера страниц, графики и рисунки, их названия, иноязычные отрывки. После того, как все тексты были сохранены в формате .docx, в программе Notepad++ во всех файлах двойные пробелы были заменены на единичные, удалены такие лишние символы как «[…]». Затем были созданы копии файлов в формате .txt с кодировкой UTF-8.

После того, как все тексты были собраны, отредактированы и сохранены в необходимом формате, мы приступили к следующему этапу составления корпуса — разметке. Как мы уже упоминали в предыдущем разделе, существуют три самых частых вида лингвистической разметки — синтаксическая, морфологическая и семантическая. Также мы выяснили, что основной из них является морфологическая разметка, то есть частеречная, она приписывает каждому токену набор граммем — род, число, падеж, одушевленность/неодушевленность, переходность, время, вид и т. д. Помимо этого, морфологическая разметка наиболее коррелирует с целью нашего корпуса — отразить в контексте основные понятия корпусной лингвистики — поэтому было решено ограничиться ей.

Наш корпус было решено составить полностью на базе корпусно-поисковой системы Sketch Engine (https://the.sketchengine.co.uk). Функционал Sketch Engine позволил нам выполнить два последних этапа одновременно, поскольку он предлагает пользователям функцию создания собственного корпуса,

автоматически размечая тексты и предоставляя полный доступ к своим функциям корпусного менеджера.

Итак, в результате выполнения этапов, описанных выше, мы получили готовый учебный корпус академических статей по корпусной лингвистике. Корпус состоит из 13 текстов (файлов), а его объем составил 1 188 505 токенов и 999 236 словоупотреблений. Из них 35 942 уникальных слова. Стоит отметить, что корпус — открытая и гибкая система, то есть его можно дополнять новыми материалами или удалять уже существующие, если они перестали быть актуальными. На рисунке 1 можно увидеть, как выглядит панель инструментов (dashboard) нашего корпуса на базе корпусно-поисковой системы Sketch Engine.

Как видно из рисунка 1, в строке поиска отображается название нашего корпуса в системе «Corpus Linguistics», а на панели инструментов можно найти такие разделы, как «Corpus Info» и «Manage Corpus». Первый раздел предоставляет нам полную информацию о корпусе, то есть его язык, список тегов, количество файлов, количество токенов, словоупотреблений, предложений, уникальных словоформ. Второй раздел дает доступ к таким инструментам, как просмотр текстов, вошедших в корпус, расширение корпуса, то есть добавление в него новых текстов, кроме этот здесь можно поделиться корпусом, скачать его, изменить настройки, добавить или отредактировать подкорпус, а также удалить корпус.

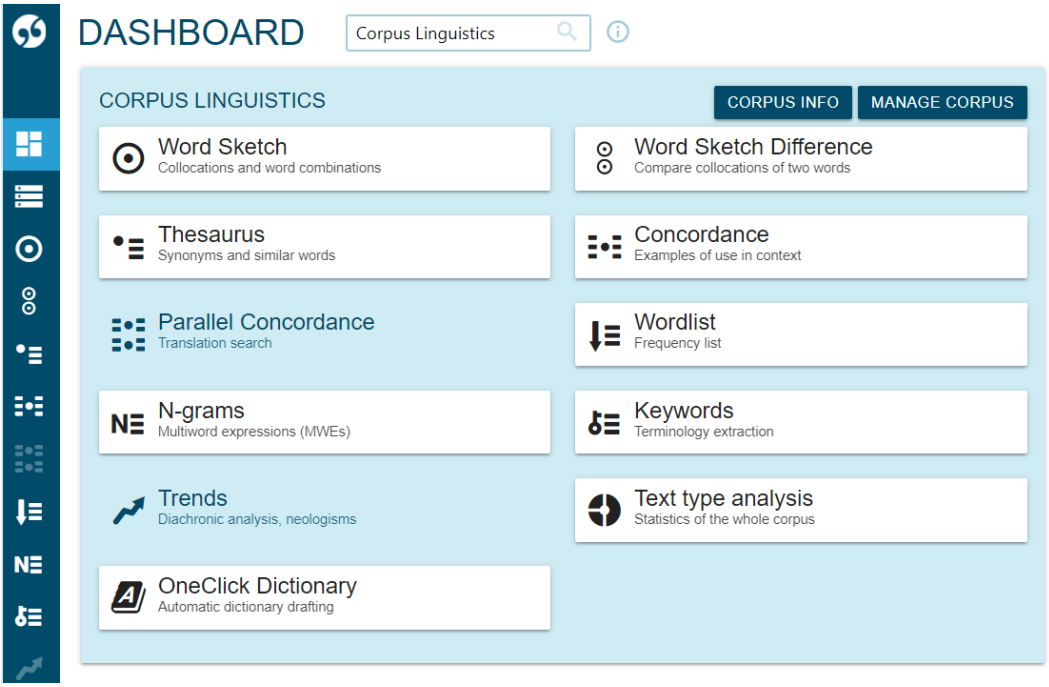

Рис. 1. Панель инструментов учебного корпуса по корпусной лингвистике.

Помимо этого, на данном рисунке мы видим все доступные нам инструменты для анализа корпуса. Таким образом, с помощью Sketch Engine мы можем строить конкордансные списки (Concordance) как по простым запросам (по слову, лемме, фразе или символу), так и по сложным запросам, составленным с помощью CQL. Кроме этого мы можем строить частотные списки слов, например существительных и глаголов, с помощью функции Wordlist, строить N-граммы (N-grams), выявлять синонимы и похожие слова для искомых терминов (Thesaurus), а также вычленять ключевые слова или терминологию, используемую в корпусе чаще всего (Keywords).

Таким образом, в данном разделе мы полностью описали процесс составления учебного корпуса академических статей по корпусной лингвистике, подробно остановившись на всех этапах: сбор материалов, редактирование, металингвистическая разметка текстов, разметка и выбор ПО для разметки текстов и работы с корпусом.

2.2. ПРОВЕРКА ФУНКЦИОНАЛЬНОСТИ УЧЕБНОГО КОРПУСА: ТЕРМИНОЛОГИЧЕСКОЕ ИССЛЕДОВАНИЕ

В данном разделе будет проведено терминологическое исследование на базе готового корпуса и представлены его результаты.

Терминологический анализ проходил в 2 этапа. На первом этапе мы провели детальный анализ ключевых терминов корпусной лингвистики.

В процессе написания теоретической части нашей выпускной квалификационной работы, в рамках которой мы описали историю, предмет и задачи корпусной лингвистики, а также практической, в которой мы определили этапы и инструменты для создания корпуса, мы дали определения ряду ключевых терминов, а именно: корпус (corpus/corpora), разметка (tagging or annotation), корпусный менеджер (corpus manager), конкорданс (concordance) и токен (token). Рассмотрим их функционирование в корпусе по порядку.

Начнем с самого главного термина для корпусной лингвистики — корпуса. Количество вхождений термина ***corpus*** составило 11 097, что составляет 0,93% от всего корпуса. Для начала определим окружение данного термина с помощью функции Word Sketch, которая определяет самые частотные коллокаты слова и группирует их по категориям грамматических отношений. Так, в категорию «определения» (modifiers) вошли такие слова, как *large, parallel, learner, comparable, small, multilingual, reference, text, general, English*. Таким образом, мы можем сделать вывод, что при описании корпусов, чаще всего говорят об их размере — 211 вхождений по запросу «large + corpus» и 70 по запросу «small + corpus» — и виде — 178 вхождений для слова parallel в сочетании с термином corpus, 139 для learner, 92 для comparable и по 59 для multilingual и reference.

Что касается существительных, используемых в подчинительной связи с термином *corpus*, самыми частотными оказались такие словосочетания, как *of corpus linguistics, corpus data, corpus linguists, corpus evidence, corpus analysis,*

corpus studies, corpus tools, corpus annotation, corpus work, research, methods. То есть, в именных словосочетаниях чаще всего используется лексика, связанная с исследованиями на базе корпуса и их результатами — *data* (417 вхождений), *evidence* (182 вхождения), *analysis* (160), *study* (132), *work* (77), *research* (74), *method* (65).

Далее рассмотрим глагольные словосочетания, в которых *corpus* выступает объектом. В данную категорию вошли глаголы *annotate, use, create, speak (spoken corpora), compile, tag (tagged corpus), specialise (specialised corpora), build, analyse*. Таким образом, чаще всего в таких сочетаниях лингвисты говорят о создании корпуса – *annotate* (165 вхождений), *create* (99), *compile* (55), *build* (38). В тех же случаях, когда корпус в сочетании с глаголом используется как субъект, чаще всего ученые говорят о нем, как о некой системе, содержащей и предоставляющей нам данные: *contain* (111 вхождений), *have* (211), *provide* (44), *include* (22), *represent* (20), *allow* (19).

Также мы составили тезаурус термина *corpus*, чтобы посмотреть похожие по значению и связанные с данным термином слова. На первом месте оказалось слово *word* (4682 вхождения), поскольку слово или словоупотребление является одной из единиц определения объема корпуса. Затем идёт слово *language* (4519 вхождений), что отсылает нас к одной из версий определения корпуса, в которой говорится, что корпус является отражением языка или языковой области. И на третьем месте расположилось слово *text* (4,043), основа, составляющая корпус. Также в контексте корпусов часто употребляются такие слова, как *datum* (2456), *study* (2332), *analysis* (2440), *example* (3082), *use* (2125) и, конечно, *linguistics* (1819).

Что касается употребления множественного числа данного термина, упомянутое во введении, вариант *corpora* встречается в корпусе 3395 раз. Вариант *corpuses* тоже существует, как мы и говорили, однако в корпусе у него всего 4 вхождения в корпус, при том что в одном из четырёх вхождений данный вариант

употреблен в контексте обсуждения правильного варианта формы множественного числа.

Далее поговорим о терминах разметки. Как мы выяснили ранее, ***tagging*** и ***annotation*** являются синонимами и оба обозначают разметку. Для того чтобы найти все словоформы для леммы *tag,* мы построили следующий CQL-запрос — [lemma="tag"]. По данному запросу было найдено 1051 вхождение, включая такие словоформы, как *tag, tagged, tagging*.

Для определения количества вхождений термина *annotation* и его глагола *annotate* мы составили CQL-запросы [lemma="annotation"] и [lemma="annotate"] соответственно. В результате для леммы *annotation* в корпусе нашлось 1367 вхождений, а для леммы *annotate* – 569, в сумме получилось 1936 вхождений, что почти в 2 раза больше, чем для термина *tagging*. Из этого можно вывод, что говоря о разметке, учёные чаще всего применяют термин *annotation*. Однако, изучив конкордансный список, построенный для леммы *tag,* мы заметили, что чаще всего данный термин применяется в контексте частеречной разметки. То есть, *annotation* – это общий термин для любой разметки, а *tagging* – преимущественно для морфологической, хотя данный термин и встречается, например, в контексте семантической разметки. Синонимичность данных терминов подтверждает и тезаурус, по запросу *tag* первым выходит глагол *annotate,* а по запросу *annotation* первым связанным с данным термином словом выходит *analysis*, а *tag* располагается на девятом месте. Функция Word Sketch выдает нам такие коллокаты к слову *tag*, как *semantically, grammatically, part-of-speach, tagged corpus, tagging program*. По запросу annotation выходят такие коллокаты, как *syntactic, linguistic, POS, corpus, morphological, semantic*. Это ещё раз подтверждает теорию о том, что аннотация является общенаучным термином для описания разметки, а теггинг — узкоспециальным термином.

Теперь проанализируем понятие «корпусный менеджер» в нашем корпусе. Построив конкорданс по термину ***corpus manager***, мы обнаружили всего 1 вхождение, что может говорить нам о том, что термин «корпусный менеджер» используется преимущественно отечественными лингвистами. Западные же учёные используют другой термин. Так, например, на официальном сайте доктора Лоуренса Энтони, при описании AntConc используется понятие *concordancer*. Данный термин встречается в корпусе уже 128 раз и, изучив все контексты, мы можем сделать вывод, что западные лингвисты используют его в значении корпусного менеджера. В тезаурусе к понятию *concordancer* можно встретить название корпусных менеджеров, как например AntConc, WebCorpLSE, WordSmith, а также названия корпусов, таких как MICASE (Michigan Corpus of Academic Spoken English) и BNCweb. Наиболее частотными коллокатами к рассматриваемому термину являются поколения корпусных менеджеров, которые были описаны в разделе 2.1 Главы 2, а именно *first-generation, second-generation, third-generation, fourth-generation*. Это подтверждает наше предположение о том, что западные лингвисты используют термин «коркондансер» (*concordancer*) в значении «корпусный менеджер».

Далее рассмотрим в корпусе термин, появившийся ещё задолго до корпусной лингвистики — конкорданс. По запросу ***concordance*** нашлось 558 вхождений, а тезаурус в первую очередь выдает слова, отображающие сущность данного понятия. Это, например, такие слова, как *list* и *line*, что указывает на форму конкорданса — список, также *result, sample, material, context, database, example, finding, datum*, поскольку конкорданс — это данные в корпусе, полученные по определённому запросу, иными словами результат, пример, контекст.

С помощью функции Word Sketch для термина *concordance* были обнаружены такие частотные коллокаты, как например *biblical*, что отсылает нас к происхождению конкордансов, описанному в Главе 1 — изначально данным

термином обозначались списки слов, основанные на Библейских текстах, с указанием стиха. Кроме этого, из коллоката *alphabetise* (in the alphabetised concordance) мы узнаем, что конкордансы могут быть расположены в алфавитном порядке.

В заключении рассмотрим термин ***токен***. Всего в учебном корпусе по корпусной лингвистике он встречается 345 раз и даже при первом взгляде на конкорданс становится видно, что наиболее частым левым коллокатом к нему будет фраза «number of». Построив список коллокатов с помощью Word Sketch, мы нашли подтверждение данному предположению — в категории «prepositional phrases» самым частотным оказался коллокат number of (tokens). А в тезаурусе вполне ожидаемо можно встретить такое слово как *hit* (вхождение), поскольку вхождения в корпус считаются в токенах. Помимо этого, в тезаурусе есть как слова *unit* (единица) и *charater* (символ), так как токен – это любая единица в тексте, будь то слово или символ, так и конкретные названия единиц, таких как *interjection* (восклицательный знак), *period* (точка), *lemma* и т. д.

На следующем этапе корпусного терминологического исследования мы выделили наиболее частотные термины.

Частотность терминов определялась при помощи инструмента Keywords, который позволяет выделять как single-words (термины, состоящие из одного слова), так и multi-word terms (термины, состоящие из двух и более слов). В связи с этим было решено выделить 100 наиболее частотных single-word terms и 100 multi-word terms, а затем распределить их на три группы: общенаучные термины, отраслевые и узкоспециальные. Под общенаучными мы понимаем такие термины, которые сохраняют свое значение, вне зависимости от области, в которой они в данный момент употребляются, под отраслевыми мы понимаем в целом лингвистические термины, а под узкоспециальными — относящиеся к корпусной лингвистике.

Для начала мы выделили топ-100 односоставных терминов, входящих в наш корпус. Вполне ожидаемо в список вошли не только термины, но и названия программ и имена учёных, их мы распределили, основываясь на их основной деятельности. Кроме этого в список вошло обозначение тега NP, сокращение *BrE*, обозначающее *British English*, аббревиатура *ICAME (International Computer Archive of Modern and Medieval English)*, которой называется международная группа лингвистов, работающих в области корпусной лингвистики для оцифровки текстов на английском языке, аббревиатура *EFL (English as a Foreign Language),* а также пары терминов *Linguistics* и *Liguistic*, *Collocate и Collocational*, которые мы посчитали как одно.

Итак, проанализировав полученные термины и проклассифицировав их описанным выше способом, мы получили данные, представленные в Таблице 3.

Таблица 3.

Топ-100 односоставных терминов в корпусе

Общенаучные	Отраслевые	Узкоспециальные
Representativeness, Discourse, Politeness, Pedagogic.	Collocation, Biber, Lexical, Linguistics, Syntactic, Linguist, Prosody, Grammatical, Collocate, Pseudo-title, Diachronic, Neo-Firthian, Semantic, Pronoun, Sociolinguistic, Noun, Infinitive, Verb, Part-of-speech, Contrastive, Adverb, Adjective, Lemma, Modal, Lexis, Adverbial, N-gram, Quotative, Lexicographer, Determiner, Preposition, Grammar, Leech, Hoey, Interjection, Co-occurence, Semi-modal, Colligation, Phraseology, Lexicology, Chomsky, Gries, Sociolinguistics, Ajimer, Pragmatic, Appositive, Stubbs, Anglicism, Louw, Demonstrative, Prosodic, Anaphora, Synchronic, Superlative, Multi-word,	Corpus, BNC, Corpus-based, Concordance, Annotation, Treebank, Corpus-driven, ICE-GB, COBUILD, McEnery, Concordancer, Hunston, Subcorpus, Subcorpora, Tagset, Tagger, Annotate, Sinclair, LOB, WordNet, FLOB, Concordancing, NP, ICAME, Taggers.

Общенаучные	Отраслевые	Узкоспециальные
	Grammarian, Phraseological, Utterance, CNCC, Semantically, BrE, Aarts, Prepositional, Coreference, Monolingual, Metalinguistic, EFL.	

Как мы видим из данных Таблицы 3, в корпусе крайне мало общенаучных терминов — порядка 4%. Основная масса (71%) терминологии приходится на отраслевую, относящуюся к лингвистике в целом, а 25% терминов — узкоспециальные, то есть относящиеся к корпусной лингвистике. Это вполне ожидаемо, поскольку корпусная лингвистика является узкой областью лингвистики, соответственно помимо терминов, связанных с корпусами, в данной области используются и общелингвистические термины.

Далее мы выделили топ-100 многосоставных терминов. В этот раз мы не обнаружили среди них имён или повторений в виде множественного числа, поэтому в таблицу (Таблица 4) вошли все сто словосочетаний.

Таблица 4.

Топ-100 многосоставных терминов в корпусе

Общенаучные	Отраслевые	Узкоспециальные
Sampling frame, information status, search interface, factor analysis, relative frequency, manual analysis, situational variation, academic writing, academic prose.	Semantic prosody, noun phrase, language use, language teaching, corpus linguistics, type noun, native speaker, linguistic analysis, semantic preference, word order, linguistic theory, relative clause, discourse presentation, credit crunch, language acquisition, conceptual metaphor, target language, linguistic research, discourse analysis, pattern grammar, lexical item, English usage, discourse marker, phrase structure, survey of English usage, text type, construction grammar, metaphor theory, linguistic description, language change, generative grammar,	Corpus evidence, corpus analysis, corpus annotation, learner corpus, corpus work, parallel corpus, corpus linguist, corpus research, syntactic annotation, linguistic annotation, annotation scheme, large corpus, corpus-driven approach, reference corpus, corpus design, corpus construction, corpus-based approach, token ratio, corpus study, corpus compiler, comparable corpus, ICE project, Sketch Engine, frequency list, diachronic corpus, multilingual corpus, using corpus, text corpus, corpus approach, word frequency.

Общенаучные	Отраслевые	Узкоспециальные
	personal pronoun, language usage, Chinese fiction, syntactic complexity, lexical priming, contrastive analysis, direct speech, running text, word class, pragmatic expression, word level, machine translation, grammatical function, conceptual metaphor theory, scientific writing, written language, speech act, child language, node word, direct object, verb phrase, information structure, press reportage, modal particle, grammatical information, language processing, lingua franca, natural language, collocatinal profile, linguistic information.	

Из данных Таблицы 4 мы видим, что многосоставных общенаучных терминов чуть больше, чем односоставных — 4 термина в Таблице 3 и 9 терминов в Таблице 4. Также в Таблице 4 больше узкоспециальных терминов — 30 единиц. В основном сюда вошли виды разметки, типы корпусов и другие термины, связанные с разработкой корпусов и исследованиями на их основе. Однако в целом ситуация довольно похожа на ту, что получилась с односоставными терминами: 9% общенаучных терминов, 61% отраслевых и 30% узкоспециальных. Из этого мы можем сделать вывод, что в научных трудах по корпусной лингвистике используются по большей части отраслевые термины, чуть больше четверти терминов — узкоспециальные (55 из 200), и всего 13 из 200 терминов — общенаучные. Полученные данные свидетельствуют о том, что корпусную лингвистику следует выделять не как подобласть компьютерной лингвистики, а как самостоятельную лингвистическую дисциплину.

Резюмируя, в данном разделе мы описали составленный учебный корпус академических статей по корпусной лингвистике, описали инструменты работы с

корпусом, а также продемонстрировали функционал созданного корпуса, проведя корпусный терминологический анализ.

ВЫВОДЫ ПО ГЛАВЕ 2

В данной главе было дано развернутое описание готового учебного корпуса по корпусной лингвистике, были подробно описаны инструменты работы с корпусом. Также были приведены результаты терминологического исследования на базе корпуса, в котором проанализировали функционирование в корпусе основных терминов корпусной лингвистики. Так, мы выяснили, что частотными коллокатами к термину *corpus* являются слова, связанные с размерами и видами корпусов, а также с исследованиями на их основе. Построили тезаурус, по данным которого узнали, что наиболее частотными контекстуальными синонимами к данному термину являются слова *word, language* и *text*.

Что касается терминов разметки, то по данным корпуса можно говорить о том, что термин *annotation* используется как общенаучный термином для описания разметки, а *tagging* чаще всего используется в контекстах морфологической разметки. Кроме этого, мы выяснили, что в англоязычных научных текстах по корпусной лингвистике не употребляют термин *corpus manager*, а вместо него используют *concordancer*. Это подтвердилось списком наиболее частотных коллокатов, на первых местах в котором номера поколений корпусных менеджеров и некоторые из их названий.

Помимо этого, в рамках данного исследования, были выделены ключевые односоставные и многосоставные термины в нашем корпусе. Было взято по 100 первых из обоих списков и распределено на три группы — общенаучные, отраслевые и узкоспециальные. В результате оказалось, что в корпусе содержится крайне мало общенаучных терминов, примерно четверть — узкоспециальных, а основную массу составили отраслевые термины.

ЗАКЛЮЧЕНИЕ

Несмотря на предпосылки, появившиеся ещё в XVIII веке, корпусная лингвистика считается молодой наукой. Именно её закрепление, как отдельной научной области, повлекло за собой возникновение множества новых терминов, не всегда до конца понятных людям, начавшим свой путь в изучении лингвистики корпусов.

Исходя из этого, настоящей задачей данного исследования являлось создание такого ресурса, который бы систематизировал труды ученых-лингвистов, специализирующихся на изучении методов и проблем корпусной лингвистики, и отразил в контексте основные понятия данной науки. Мы посчитали, что самым лучшим решением этой задачи будет составление учебного корпуса по корпусной лингвистике, поскольку это наиболее удобный в использовании формат, предоставляющий пользователю множество статистических данных по каждому элементу текста.

В работе было предоставлено теоретическое обоснование корпусной лингвистики. В первую очередь, были выделены этапы её развития как науки, начиная с появления первых конкордансов в XVIII веке, заканчивая веб-корпус-менеджерами четвёртого поколения, позволяющими пользователям создавать большие корпуса для лингвистических исследований любой тематики за счёт хранения данных на серверах.

Далее мы провели сравнительно-сопоставительный анализ трактовок прикладной лингвистики отечественных и зарубежных учёных и пришли к выводу, что отечественные лингвисты связывают прикладную лингвистику в первую очередь с разработкой компьютерных программ обработки языка, составлением корпусов и исследованиями на их основе, а зарубежные ученые первостепенными задачами данной науки считают обучение языку. Помимо этого мы выяснили, что среди лингвистов нет единого мнения, к какой области относить корпусную

лингвистику — к прикладной или к компьютерной лингвистике. Проанализировав их мнения, мы пришли к выводу, что корпусная лингвистика является подобластью компьютерной, которая, в свою очередь, является подобластью прикладной. Однако, терминологическое исследование на базе нашего корпуса показало, что корпусную лингвистику и вовсе можно выделить как отдельную лингвистическую область, поскольку в связанных с ней работах преобладает отраслевая терминология.

Помимо этого было изучено и центральное понятие корпусной лингвистики — непосредственно лингвистический корпус текстов. Некоторые учёные считают корпус уменьшенной моделью языка, поскольку достаточно большой и репрезентативный корпус, которыми, например, являются национальные корпуса, действительно отражают функционирование языка в различных областях и, что немаловажно, отражают не только литературный язык, но и живой, то есть «узус, а не норму». Также мы выяснили, что лингвистический корпус отличается от простого собрания текстов по определённой теме наличием лингвистической аннотации, или разметки.

Таким образом, изучив само понятие корпусной лингвистики и корпуса, мы приступили к сбору и систематизации информации о составлении корпусов. В рамках этого мы выделили 4 основных этапа создания корпуса — разработка проекта, сбор, оцифровка и редактирование материалов, разметка и выбор корпусного менеджера. Каждый этап был подробно описан в разделе 1.3.1, была обозначена цель каждого этапа, составлен алгоритм и способы его выполнения, представлено программное обеспечение, необходимое для более быстрой и точной работы, обозначено, что должно получиться в результате каждого этапа.

Затем, опираясь на собранные данные, мы составили учебный корпус академических статей по корпусной лингвистике, объёмом в 1 188 505 токенов, подробно описав каждый этап нашей работы. Так, была выполнена

экстралингвистическая разметка всех материалов, вошедших в корпус, описан процесс оцифровки и редактирования отобранных текстов, обоснован выбор использованного программного обеспечения.

На базе полученного корпуса мы провели два вида терминологического исследования. Первый заключался в том, что мы исследовали окружение и функционирование в корпусе таких понятий корпусной лингвистики, выделенных в теоретической части работы, как корпус, разметка, корпусный менеджер, токен и конкорданс.

Так, из результатов исследования понятия corpus хотелось бы выделить, что при описании корпуса лингвисты в своих работах чаще всего говорят о его размерах и видах. В тезаурусе же были выделены наиболее частые контекстуальные синонимы и связанные с данным понятием слова, ими оказались *word, language* и *text,* что в целом прекрасно описывает понятие корпуса.

Исследование понятия разметки на базе корпуса позволило провести более четкую разницу между терминами *annotation* и *tagging*. Изучив конкордансы по запросу *tagging,* мы заметили, что данный термин чаще всего употребляется в контексте морфологической разметки. Таким образом, корпусное исследование показало, что annotation – общий термин для понятия разметки, включающий в себя как морфологическую, так и другие виды разметки.

Что касается термина «корпусный менеджер», в результате корпусного исследования мы выяснили, что в англоязычных работах, посвященных корпусной лингвистике, не употребляется прямой перевод — *corpus manager*. Вместо этого, англоязычный термин *concordancer* отражает основную функцию корпусного менеджера — построение конкордансов. Это подтверждается и списком частотных коллокатов, возглавляют который номера поколений корпусных менеджеров.

Сам же термин «конкорданс» в тезаурусе, построенном на базе корпуса, смежными понятиями имеет такие слова как *list* и *line,* что наиболее точно

отражает его форму. Наиболее же частотный коллокат — *biblical* — отсылает нас к первоначальному определению данного термина и его возникновению в контексте священных писаний.

Говоря о наиболее маленькой единице корпуса — токене, корпусное исследование подтверждает её функционирование как единицы измерения объёма корпуса. Так, наиболее частотным коллокатом к термину *token* является фраза *number of*. А в тезаурусе можно встретить такие слова, как *unit* и *character*, поскольку токен — это любая единица текста, будь то слово, лемма или символ.

В рамках второго терминологического исследования были выделены наиболее частотные термины в корпусе — 100 односоставных (single-words) и 100 многосоставных (multi-word terms). Выделенные термины затем мы разделили на три группы — общенаучные, отраслевые и узкоспециальные. В результате данного исследования было выявлено, что в научных текстах по корпусной лингвистике в основном используются отраслевые термины, то есть термины, относящиеся к лингвистике в целом, чуть более четверти от общего числа составляют узкоспециальные, относящиеся к корпусной лингвистике, и лишь малая часть (11 из 200) приходится на общенаучные термины.

Говоря о практическом применении данной работы, хотелось бы сказать, что составленный учебный корпус по корпусной лингвистике может быть использован как в учебных целях, то есть с его помощью него можно глубже разобраться в терминологии корпусной лингвистики, так и в научных, поскольку на его основе можно проводить любые корпусные исследования. Помимо этого, в данной работе представлено подробное пошаговое описание процесса создания корпуса, так что ее можно использовать как инструкцию по созданию корпуса по любой тематике и для любых целей.

Важно также упомянуть перспективы данного исследования. Поскольку наш корпус является открытой системой, он может быть дополнен и расширен. Причем

возможно как добавление новых материалов по корпусной лингвистике, так и возможно расширять область, охваченную данным корпусом, например в сторону компьютерной лингвистики или лингвистики в целом.

Исходя из всего вышесказанного, можно заключить, что цель, поставленная во введении данной работы, достигнута, а задачи выполнены в полном объёме.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Базарова Б. Б. Введение в корпусную лингвистику: учебно-методическое пособие. Улан-Удэ: Издательство Бурятского госуниверситета, 2016. 76 с.

2. Баранов А. Н. Введение в прикладную лингвистику: Учебное пособие. М.: Эдиториал УРСС, 2001. 360 с.

3. Большакова Е. И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учебное пособие. М.:МИЭМ, 2011. 272с.

4. Боярский К. К. Введение в компьютерную лингвистику. Учебное пособие. Спб: НИУ ИТМО, 2013. 72 с.

5. Волоснова Ю. А. Корпусная лингвистика: проблемы и перспективы. Лесной вестник, 2006. 43-49 с.

6. Ганиева И. Ф. Об использовании корпусов в лингвистических исследованиях. Вестник Башкирского университета, 2007. 104-106 с.

7. Грудева Е. В. Корпусная лингвистика: учеб. Пособие. М. : ФЛИНТА, 2012. 165 с.

8. Захаров В. П., Азарова И. В. Модель программно-лингвистического комплекса для создания и использования специализированных корпусов русского языка. СПб.: Изд-во С.-Петерб. Ун-та, 2018. 208 с.

9. Захаров В. П., Богданова С. Ю. Корпусная лингвистика: Учебник для студентов гуманитарных вузов. Иркутск: ИГЛУ, 2011. 161 с.

10. Захаров В. П. Корпусная лингвистика в России. Доклад на конференции: IV международный научный симпозиум Retro'2013. Ретроспектива филологии в информационном обществе знаний, 2013. 9 с.

11. Захаров В. П. Корпусная лингвистика: Учебно-метод. Пособие. СПб., 2005. 48 с.

12. 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): В 2 т. Т. 1: Основная программа конференции. М.: Изд-во РГГУ, 2015. С. 667-682.

13. Зубов А. В. Информационные технологии в лингвистике: Учеб. пособие для студ. лингв, фак-тов высш. учеб. Заведений. М.: Издательский центр «Академия», 2004. 208 с.

14. Зубов А. В. Корпусная лингвистика: возможности и перспективы. Материалы пленарного заседания, 2006. 22-27 с.

15. Инструментарий естественного языка — Natural Language Toolkit [Электронный ресурс]. URL: https://ru.qaz.wiki/wiki/Natural_Language_Toolkit (дата обращения 21.05.2020)

16. Калинина Е. История и основные понятия корпусной лингвистики // Изучаем Digital Humanities [Электронный ресурс]. 2018. URL: https://dhumanities.ru/?p=667 (дата обращения: 19.05.2020).

17. Козлова Н. В. Лингвистичекие корпуса: определение основных понятий и типология. Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация, 2013. Том 11, выпуск 1. 79-88 с.

18. Копотев М. В. Введение в корпусную лингвистику. [Электронный ресурс]. Прага: Animedia Company, 2014. 218 с. URL: http://animedia-company.cz/ebookscatalog/business-popular-science-catalog/vvedenije-v-korpusnuju-lingvistiku (дата обращения 20.11.2019)

19. Корпусно-поисковая система Sketch Engine [Электронный ресурс]. URL: https://www.sketchengine.eu/ (дата обращения 03.06.2021)

20. Корпусы и корпусная лингвистика. Основные понятия. — Текст : электронный // Myfilology.ru – информационный филологический ресурс : [сайт]. – URL: https://myfilology.ru//177/korpusy-i-korpusnaya-lingvistika-osnovnye-ponyatiya/ (дата обращения: 31.05.2021)

21. Кутузов А. Б. Корпусная лингвистика: курс лекций. ТюмГУ. 45 с.

22. Лаврентьев А. М. Корпусная лингвистика: идеология, методы, технологии. Сибирский филологический журнал, 2004. №3-4. 121-134 с.

23. Майорова А. Д. Корпусная лингвистика: исторический и лингводидактический аспекты. Международный научно-исследовательский журнал, 2017. № 05 (59), Часть 2. 42-46 с.

24. Марчук Ю. Н. Компьютерная лингвистика: учебное пособие. М.: АСТ: Восток — Запад, 2007. 317 с.

25. Национальный корпус русского языка [Электронный ресурс]. URL: http://www.ruscorpora.ru/new/corpora-about.html#task (дата обращения 04.10.2020)

26. Николаев И. С., Митренина О. В., Ландо Т. М. (Ред.). Прикладная и компьютерная лингвистика. М.: URSS, 2016. 320 с.

27. Овчинникова И. Е., Угланова И. А. Компьютерное моделирование вербальной коммуникации: учеб.-метод, пособие. М.: Флинта: Наука, 2009. 136с.

28. Плунгян В. А. Как создаются и используются корпуса языков. [Электронный ресурс]. Постнаука, 2013 URL: https://postnauka.ru/faq/87397 (дата обращения 15.03.2020)

29. Российская академия наук. Институт проблем передачи информации им. А. А. Харкевича [Электронный ресурс]. URL: http://iitp.ru/ru/science/works/452.htm (дата обращения 08.10.2020)

30. Савчук С. О. Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. — М., 2005, 62—88

31. Семенова Э. В. Основы теоретической и прикладной лингвистики. Улан-Удэ: Издательство Бурятского госуниверситета, 2015. 64 с.

32. Склярова Н. Г. Введение в прикладную лингвистику. Информационные технологии в лингвистике: учебное пособие. Пятигорск: Изд-во ПГУ, 2016. 86 с.

33. Соснина Е. П. Введение в прикладную лингвистику: учебное пособие. Ульяновск: УлГТУ, 2000. 46 с.

34. Щипицина Л. Ю. Информационные технологии в лингвистике : учеб. пособие М. : ФЛИНТА : Наука, 2013. 128 с.

35. Гвишиани Н. Б. Практикум по корпусной лингвистике: Учеб. пособие на англ. яз. М., Высшая школа, 2008. 191 с.

36. Baker P. Contemporary corpus linguistics. London: Continuum, 2009. 368 p.

37. Baker P., Hardie A., McEnery T. A Glossary of Corpus Linguistics. Edinburgh University Press, 2006. 192 p.

38. Baker P. Using corpora in discourse analysis. A&C Black, 2006. 206 p.

39. Baker P. Using Corpora to Analyze Gender. Bloomsbury Academic, 2014. 235 p.

40. Caws C., Hamel M.-J. Language-Learner Computer Interactions Theory, methodology and CALL applications. John Benjamins B.V., 2016. 275 p.

41. Cook G. Applied Linguistics. Oxford University Press, 2003. 72 p.

42. Cook V., Wei L. Contemporary applied linguistics. Continuum International Publishing Group, 2009. 288 p.

43. Everaert M. The use of Databases in Cross-Linguistic Studies. Mouton de Gruyter Berlin – New York, 2009. 415 p.

44. Facchinetti R. Corpus Linguistics 25 Years on. Editions Rodopi B.V., Amsterdam – New York, 2007. 392 p.

45. Finegan E. LANGUAGE: its structure and use. N.Y.: Harcourt Brace College Publishers, 2004. 592p.

46. Flowerdew L. Corpora and language education. Palgrave Macmillan UK, 2012. 358 p.

47. Francis W. N. Problems of Assembling and Computerizing Large Corpora. In: «Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora». Königstein, 1979. Pp. 110-123.

48. Garside R., Leech G., McEnery T. Corpus annotation: linguistic information from computer text corpora. Routledge, 2016. 292 p.

49. Granger S., Hung J., Petch-Tyson S. Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. John Benjamins B.V., 2002. 258 p.

50. Kawaguchi Y., Takagaki T. Corpus-Based Perspectives in Linguistics. John Benjamins B.V., 2007. 450 p.

51. Kilgarriff A., Baisa V., Busta J., Jakubicek M., Kovar V., Michelfeit J., Rychly P., Suchomel V. The Sketch Engine: ten years on [Электронный ресурс] // Lexicography ASIALEX. 2014. V. 1. Pp.7-36. URL: http://link.springer.com/article/10.1007/s40607-014-0009-9 (дата обращения 21.05.2021)

52. Kuebler S., Zinsmeister H. Corpus linguistics and linguistically annotated corpora. Bloomsbury Academic, 2015. 321 p.

53. McEnery T., Wilson A. Corpus linguistics. Edinburgh University Press, 2001. 122 p.

54. Meyer C. F. English corpus linguistics. Cambridge University Press, 2004. 185 p.

55. Paltridge B., Phakiti A. Research Methods in Applied Linguistics. Bloomsbury Academic, 2015. 451 p.

56. Sinclair J. Corpus, concordance, collocation. Oxford University Press, 1991. 179 p.

57. Sinclair McH. J. How to Use Corpora in Language Teaching. John Benjamins B.V., 2004. 317 p.

58. Teubert W. Text corpora and multilingual lexicography. John Benjamins Publishing Company, 2007. 172 p.

59. Thomas J. Discovering English with Sketch Engine: A Corpus-Based Approach to Language Exploration, 2016. 228p.

60. TreeTagger – a part-of-speech tagger for many languages. [Электронный ресурс]. URL: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (дата обращения 05.04.2021)

61. Wołk K. Machine Learning in Translation Corpora Processing. CRC Press, 2019. 281 p.

ПРИЛОЖЕНИЕ 1

Список тегов, используемых Sketch Engine

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, one
CDZ	possesive pronoun	one’s
DT	determiner	the
EX	existential there	there is
FW	foreign word	d’hoevre
IN	preposition, subordinating conjunction	in, of, like
IN/that	that as subordinator	that
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NNSZ	possessive noun plural	people’s, women’s
NNZ	possessive noun, singular or mass	year’s, world’s
NP	proper noun, singular	John
NPS	proper noun, plural	Vikings
NPSZ	possesive proper noun, plural	Boys’, Workers’
NPZ	possesive noun, singular	Britain’s, God’s
PDT	predeterminer	both the boys
PP	personal pronoun	I, he, it

POS Tag	Description	Example
PPZ	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give up
SENT	Sentence-break punctuation	. ! ?
SYM	Symbol	/ [= *
TO	infinitive ‘to’	togo
UH	interjection	uhhuhhuhh
VB	verb be, base form	be
VBD	verb be, past tense	was, were
VBG	verb be, gerund/present participle	being
VBN	verb be, past participle	been
VBP	verb be, present, non-3d person	am, are
VBZ	verb be, 3rd person sing. present	is
VH	verb have, base form	have
VHD	verb have, past tense	had
VHG	verb have, gerund/present participle	having
VHN	verb have, past participle	had
VHP	verb have, sing. present, non-3d	have
VHZ	verb have, 3rd	has

POS Tag	Description	Example
	person sing. present	
VV	verb, base form	take
VVD	verb, past tense	took
VVG	verb, gerund/present participle	taking
VVN	verb, past participle	taken
VVP	verb, present, not 3rd person	take
VVZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WPZ	possessive wh-pronoun	whose
WRB	wh-abverb	where, when
Z	possessive ending	‘s

ПРИЛОЖЕНИЕ 2

Метаразметка текстов, вошедших в корпус

Информация об авторе	**Информация о тексте**	**Информация об аудитории**	**Библиографическое описание текста**
• Paul Baker	• Using Corpora	• Н-возраст	• LexisNexis, a

Информация об авторе	Информация о тексте	Информация об аудитории	Библиографическое описание текста
• мужской • 1972г.	in Discourse Analysis • 2006 • 118 908 словоупотреблений • учебно-научная сфера • Корпусная лингвистика • монография • научный стиль	• профессиональный уровень	division of Reed Elsevier Inc. • Электронная версия
• Roberta Faccinetti • женский • 1967г.	• Corpus Linguistics 25 Years on • 2007 • 112 771 • учебно-научная сфера • корпусная лингвистика • монография • научный стиль	• Н-возраст • профессиональный уровень	• Editions Rodopi B.V., Amsterdam – New York • Электронная версия
• Eileen Fitzpatrick • женский	• Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse • 2007 • 85 096 словоупотреблений (отрывок) • учебно-научная сфера • корпусная лингвистика • монография • научный стиль	• Н-возраст • профессиональный уровень	• Editions Rodopi B.V., Amsterdam – New York • Электронная версия
• Lynne	• Corpora and	• Н-возраст	• Palgrave

Информация об авторе	Информация о тексте	Информация об аудитории	Библиографическое описание текста
Flowerdew	Language Education • 2012 • 94 791 • учебно-научная сфера • корпусная лингвистика • монография • научный стиль	• профессиональный уровень	Macmillan • электронная версия
• Gvishiani Natalia Borisovna • женский	• English on Computer: a Tutorial in Corpus Linguistics • 2008 • 35 559 • учебно-научная сфера • корпусная лингвистика • учебное пособие • научный стиль	• Н-возраст • высокий уровень	• ОАО «Издательство «Высшая школа» • электронная версия
• Sandra Kübler, Heike Zinsmeister	• Corpus Linguistics and Linguistically Annotated Corpora • 2015 • 75 212 • учебно-научная сфера • корпусная лингвистика • монография • научный стиль	• Н-возраст • профессиональный уровень	•Bloomsbury Publishing Plc •электронная версия
• Anke Lüdeling, Merja Kytö • женский • 1968г.,	• Corpus Linguistics An International Handbook • 2008	• Н-возраст • профессиональный уровень	• Walter de Gruyter • электронная версия

Информация об авторе	Информация о тексте	Информация об аудитории	Библиографическое описание текста
1953г.	• 57 094 • учебно-научная сфера • корпусная лингвистика • монография • научный стиль		
• Tony McEnery, Andrew Hardie • мужской	• Corpus Linguistics: Method, Theory and Practice • 2012 • 120 931 • учебно-научная сфера • корпусная лингвистика • монография • научный стиль	• Н-возраст • профессиональный уровень	• Cambridge University Press • электронная версия
• Charles F. Meyer • мужской	• English Corpus Linguistics: an Introduction • 2004 • 63 684 • учебно-научная сфера • корпусная лингвистика • монография • научный стиль	• Н-возраст • профессиональный уровень	• Cambridge University Press • электронная версия
• Antoinette Renouf, Andrew Kehoe	• Corpus Linguistics: Refinements and Reassessments • 2009 • 148 005 • учебно-научная сфера • корпусная лингвистика	• Н-возраст • профессиональный уровень	• Editions Rodopi B.V., Amsterdam – New York • Электронная версия

Информация об авторе	Информация о тексте	Информация об аудитории	Библиографическое описание текста
	• сборник статей • научный стиль		
• James Thomas • мужской	• Sketch Engine: a Toolbox for Linguistic Discovery • 2016 • 1814 • учебно-научная сфера • корпусная лингвистика • статья • научный стиль	• Н-возраст • профессиональный уровень	• Discovering English with Sketch Engine: a Corpus-Based Approach to Language Exploration. 2nd ed. Brno: Versatile, 2016. • электронная версия
• Elena Tognini-Bonelli • женский	• Corpus Linguistics At Work • 2001 • 74 440 • учебно-научная сфера • корпусная лингвистика • монография • научный стиль	• Н-возраст • профессиональный уровень	• John Benjamins B.V. • электронная версия
• Wolfgang Teubert • мужской	• Text Corpora and Multilingual Lexicography • 2007 • 39 926 • учебно-научная сфера • корпусная лингвистика • сборник статей • научный стиль	• Н-возраст • профессиональный уровень	• John Benjamins B.V. • электронная версия