МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САМАРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА»
(САМАРСКИЙ УНИВЕРСИТЕТ)

Т.В. ТОЛСТОВА

ЖАНР И КОРПУС: СОВРЕМЕННЫЕ ПОДХОДЫ К ИЗУЧЕНИЮ И ПРЕПОДАВАНИЮ ЯЗЫКА

Одобрено редакционно-издательским советом федерального государственного автономного образовательного учреждения высшего образования «Самарский национальный исследовательский университет имени академика С.П. Королева»

САМАРА
Издательство Самарского университета
2018

ББК 81.2
УДК 81
Т 529

Рецензенты: д-р филол. наук, проф. М. М. Х а л и к о в,
д-р филол. наук В. Д. Ш е в ч е н к о

Толстова Татьяна Витальевна

Т529 **Жанр и корпус: современные подходы к изучению и преподаванию языка** / *Т.В. Толстова*. – Самара: Изд-во Самарского университета, 2018. – 196 с.: ил.

ISBN 978-5-7883-1324-5

В монографии предпринимается попытка синтеза двух перспективных направлений в науке о языке – жанрового анализа и корпусной лингвистики. Рассматриваются различные аспекты данных направлений в процессе изучения функционирования языка и в обучении продуцирования речи различных жанров.

Анализируются основной инструментарий и возможности корпусной лингвистики – направления прикладной лингвистики, использущего достижения компьютерных технологий в обработке и исследовании больших электронных массивов языкового материала

Адресуется широкому кругу специалистов в области лингвистики и методики обучения иностранным языкам, занимающихся научными исследованиями в области языкознания и разработкой учебных и методических материалов.

Кафедра иностранных языков и русского как иностранного.

УДК 81
ББК 81.2

ISBN 978-5-7883-1324-5

Оглавление

Введение

Научно-техническая и информационная революции XX века, охватившие все области человеческого знания и практики, не могли не затронуть и языкознания. Речь идет не только о новых формах фиксации, таких как аудио– или видеозапись и электронное хранение данных, а о явлениях, принципиально меняющих наши взгляды и подходы к функционированию языка. В последние годы широкое распространение получила корпусная лингвистика – направление прикладной лингвистики, целью которого является создание и обслуживание (с использованием машинных носителей и соответствующих средств доступа) больших массивов языковых данных.

В связи с широким распространением и применением корпусной лингвистики неизбежно возникает вопрос, является ли она самостоятельной отраслью лингвистики, каково ее место и роль в иерархии языковедческих дисциплин? По поводу этого вопроса существуют различные взгляды и мнения. Даже если корпусная лингвистика еще не стала отдельным лингвистическим направлением, а выступает лишь в роли особой технологии, то технологии поистине «революционной». По словам Дж. Суэйлза (Simpson, Swales 2001) «как телескоп в свое время изменил астрономию, рентген революционизировал медицину, магнитофон подтолкнул развитие социолингвистики, видеомагнитофон продвинул вперед изучение взаимоотношений в малых социальных группах, а спектрограф консолидировал развитие инструментальной фонетики – так и корпусная лингвистика не могла не привести к радикальным результатам».

Сегодня корпусная лингвистика активно и успешно применяется в различных областях, таких как прикладная лингвистика, филология,

стилистика, методика обучения иностранным языкам, теория и практика перевода, криминальная лингвистика, контент-анализ, анализ дискурса, жанровый анализ.

Говоря о жанровом анализе, нельзя не отметить, что владение теми или иными жанрами речи помогает людям успешно общаться в различных ситуациях социального и делового общения. Изучение законов, или конвенций, жанра является еще одним важнейшим направлением в современном речеведении.

В данной работе предпринимается попытка синтеза двух перспективных подходов к изучению и преподаванию текстов. С этой целью в первой главе рассматриваются три взаимосвязанных направления в исследовании текста – лингвистика текста, дискурс, жанр. В качестве основной особенности жанрового подхода выделяется анализ когнитивной структуры текста, выполняемый с помощью выделения коммуникативных шагов, из которых складывается единство текста.

В последующих главах детально представлены сущность и возможности корпусной лингвистики: начиная с того, является ли она отдельной дисциплиной или лишь инструментом, до типологии корпусов и их использования для разных видов лингвистических исследований (грамматических, лексических и жанровых), а также в обучении иностранным языкам.

1. Лингвистика текста – дискурс – жанр: эволюция и взаимопроникновение понятий

Речевые произведения различного объема, различной функциональной и стилистической направленности сегодня привлекают к себе пристальное внимание исследователей и преподавателей языков в силу того, что именно знание законов построения таких произведений и умение продуцировать их быстро, естественно и наиболее уместным и оправданным для конкретной ситуации способом свидетельствует о полноценном владении языком как средством общения.

В связи с природой и сущностью таких произведений в последнее время в лингвистической литературе можно наблюдать некоторое пересечение, а подчас дублирование нескольких понятий. Вызвано это, прежде всего, различиями в научных школах, подходах и точках зрения. В то же время такое пересечение свидетельствует об актуальности и важности этих понятий. Речь идет о лингвистике текста, дискурс-анализе и жанровом анализе.

Мы не ставим перед собой цели дать исчерпывающий теоретический обзор данных сложных и глубоких понятий, задача состоит в том, чтобы выяснить, как они могут помочь в процессе анализа имеющихся типичных произведений речи, чтобы обеспечить впоследствии возможность строить по их образцу собственные, соответствующие языковым нормам, принятым в той или иной культуре. Рассмотрим каждое из них более подробно с точки зрения их развития и взаимосвязи между ними.

§ 1.1 Лингвистика текста: подходы и категории

Роль лингвистики текста[1] обусловлена стремлением глубже изучить связи языка с различными сторонами человеческой деятельности, реализуемыми через текст. Исследования, проводимые как в нашей стране, так и за рубежом, позволяют проводить детальные разработки сущностных характеристик общей лингвистики текста. Дискуссионным является вопрос о границах лингвистики текста, а именно: включает ли она в себя прагматику, фукциональную семантику и синтаксис, активно развивающуюся в настоящее время риторику или пересекается с ними как пересекается с поэтикой, психолингвистикой и теорией коммуникации, входя как составная часть в теорию текста.

На первом этапе своего развития в 60-е годы XX в. лингвистика текста в основном изучала способы сохранения связанности и понятности текста, методы передачи кореференции лица и предмета (анафорические структуры, лексические повторы, видовременные цепочки и т.д.), распределение темы и ремы высказывания в соответствии с требованиями актуального членения предложения. Успешному развитию этих исследований способствовали ранние работы по анафорико-катафорическим структурам, порядку слов, правилам выбора актуализации при переходе от языка к речи (В.В. Виноградов, В. Матезиус, З. Харрис и др.).

Однако поиски средств только формальной связанности текста привели к некоторому повторению тематики, отсутствию теоретических обобщений и невозможности выявления содержательных, а не формальных категорий. Лингвистика текста определяет смысловые различия, употребление коммуникативно-ориентированных компонентов высказывания – артиклей, модально-коммуникативных частиц, оценочных прилагательных, видов глагола, акцентных подчеркиваний и т.п. Выявляемые при этом смысловые различия относятся как к правилам логического развертывания

[1] Лингвистика текста представляет собой «направление лингвистических исследований, объектом которого являются правила построения связного текста и его смысловые категории, выражаемые по этим правилам» (Николаева Т.М. Лингвистика текста // ЛЭС / Гл. ред. В.Н. Ярцева. – М., 1990. – С. 267.).

содержания текста, так и к правилам прагматического характера, определяющим некоторый общий фонд знаний, общую для автора и воспринимающего «картину мира», без единства которой текст будет непонятен. Это относится к так называемым пресуппозициям. Под текстом в данном случае понимается широкое контекстно-конситуативное коммуникативное окружение – существующее, подразумеваемое, или создаваемое автором при желании воздействовать на воспринимающего.

Первое время наряду с наименованием «лингвистика текста» употреблялись «грамматика текста», «теория текста» и др. Провести строгое различие между этими названиями трудно, но представляется возможным различать четыре основных направления исследований, которые рассматриваются разными авторами либо по отдельности, либо вместе.

Во-первых, это грамматика текста, изучающая характер преобразований предложения в зависимости от его контекстных связей. Особенность этой грамматики заключается в более полном представлении о ряде традиций грамматических категорий; в появлении новых содержательных категорий; выявлении новых функциональных классов; переосмыслении новой классификации ряда единиц традиционной грамматики.[2]

Во-вторых, это собственно лингвистика текста – исследование текста на всех уровнях: семантики, синтактики, прагматики. Большинство работ в этих двух направлениях[3] можно охарактеризовать как своеобразное «продолжение» традиционной грамматики, вышедшей за пределы одного предложения. В связи с этим проводится ряд аналогий между структурами сложного предложения и текста.

В-третьих, это интерпретация текста, которая стала предметом преподавания на филологических факультетах. И, в-четвертых, проблема типологии текстов – попытки выявить параметры и

[2] См. об этом подробнее: Николаева Т.М. Лингвистика текста: Современное состояние и перспективы // НЗЛ. М., 1978. – Вып. 8.

[3] См. в связи с этим: Папина Ф.А. Текст: его единицы и глобальные категории. – М.: Едиториал УРСС, 2002. – С. 13–15.

критерии, которые позволили бы осуществить классификацию всех имеющихся текстов.

Один из первых исследователей лингвистики текста немецкий ученый П. Хартман[4] предлагал разделить сферы исследования текста на общую лингвистику текста, лингвистику конкретного текста и лингвистику типологии текстов. В дальнейших исследованиях наблюдается два подхода:[5]

- стремление построить формализованную грамматику текста, для чего создаются правила, процедуры, схемы, по которым можно осуществить моделирование текста;
- стремление создать общую теорию текста путем изучения конкретных речетворческих актов, закономерностей их организации и функционирования, описание стилевого многообразия таких актов и определение категориальных признаков каждого типа текстов.

Первый подход характерен для западноевропейских школ, а второй подход – для отечественной лингвистики.

З.Я. Тураева выделяет несколько направлений, по которым развивается лингвистика текста:[6]

- изучение текста как системы высшего ранга, основными признаками которой являются целостность и связность;
- построение типологии текстов по коммуникативным параметрам и соотнесеным с ними лингвистическим признакам (понимаемым широко в единстве плана выражения и плана содержания);
- изучение единиц, составляющих текст;
- выявление особых текстовых категорий;
- определение качественного своеобразия функционирования языковых единиц различных уровней под влиянием текста, в результате их интеграции с текстом;
- изучение межфразовых связей и отношений.

---

[4] См. об этом: Гальперин И.Р. Текст как объект лингвистического исследования. – М.: Едиториал УРСС, 2004. – С. 8.

[5] Гальперин И.Р., 2004. – С 8.

[6] Тураева З.Я. Лингвистика текста (текст: структура и семантика). – М.: Просвещение, 1986. – С. 7–8.

Ю.А. Левицкий в работе «Лингвистика текста» (2006) выделяет три аспекта, с помощью которых может быть изучен текст: прагматический, психолингвистический и социолингвистический. В свою очередь, З.Я. Тураева (1986) рассматривает различные аспекты самого текста:

- онтологический аспект – характер существования текста, его статус, отличие от устной речи;
- гносеологический аспект – характер отображения объективной действительности в тексте, а в случае художественного текста – характер отражения реального мира в идеальном мире эстетической действительности;
- собственно лингвистический аспект – характер языкового оформления текста;
- психологический аспект – характер восприятия текста;
- прагматический аспект – характер отношения автора текста к объективной действительности и к содержательному материалу.

Объектом изучения лингвистики текста является текст, в отношении которого существуют различные концепции (Тураева 1986), в которых данное понятие интерпретируется в зависимости от того, какой аспект текста в них выделяется как ведущий:

1. Концепции, в которых ведущим считается статистический аспект, можно объединить как отражающие результативно-статическое представление о тексте. Текст понимается как информация, отчужденная от отправителя.

2. Концепции, в которых на первый план выдвигается процессуальность текста как реализация речевой способности человека, с одной стороны, и как способность языка к живому функционированию в речи – другой.

3. Концепции, акцентирующие каузирующее начало – речевую деятельность индивидуума как источник текста –ориентируются на акт коммуникации, который предполагает наличие отправителя и получателя.

4. Стратификационные концепции, рассматривающие текст как уровень языковой системы, где включение текста в иерархию языковых уровней предполагает рассмотрение некоего абстрактного

текста (алгоритма его порождения, моделей, схем) и текста в конкретной реализации.

Классическое определение текста принадлежит И.Р. Гальперину (2004:18)[7] «Текст – это произведение речетворческого процесса, обладающее завершенностью, объективированное в виде письменного документа, литературно обработанное в соответствии с типом данного документа, произведение, состоящее из названия (заголовка) и ряда особых единиц (сверхфразовых единств), объединенных разными типами лексической, грамматической, логической, стилистической связи, имеющее определенную целенаправленность и прагматическую установку».

Текст как факт речевого акта системен. Он представляет собой некое завершенное сообщение, обладающее своим содержанием, организованное по абстрактной модели одной из существующих в литературном языке форм сообщений и характеризуемое своими отличительными признаками. Различные типы текстов могут содержать несколько видов информации: содержательно-фактуальная информация (сообщение о фактах, событиях, процессах, происходящих, происходивших, которые будут происходить в окружающее мире, действительном или воображаемом), содержательно-концептуальная информация (индивидуально-авторское понимание отношений между явлениями, понимание их причинно-следственных связей и значимости), а также содержательно-подтекстовая информация (скрытая информация, извлекаемая благодаря способности единиц текста порождать ассоциативные и коннотативные значения.

Категории текста носят характер универсалий и обнаруживаются в связном тексте независимо от языка, на котором создан данный текст, и независимо от типа текста. Коммуникативный подход к членению речевого потока предполагает выделение единиц текста по принципу их коммуникативной целостности, т.е. актуальное членение

[7] См. также определение З.Я. Тураевой (1986:11): «Текст – это некое упорядоченное множество предложений, объединенных различными типами лексической и грамматической связи, способное представить определенным образом организованную и направленную информацию. Текст есть сложное целое, функционирующее как структурно-семантическое единство».

высказывания. В процессе изучения актуального членения предложения учеными были выделены диремные и моноремные структуры, определены понятия тематической, рематической и диффузных зон высказывания, вторичной рематизации и т.д. (Л.В. Щерба, И.И. Ковтунова, Т.М. Николаева, А.П. Сковородников). Два компонента актуального членения предложения образуют отношение, аналогичное отношению между логическими субъектом и предикатом, одной стороны, и грамматическими подлежащим и сказуемом – с другой.

Одним из направлений лингвистики текста является выявление единиц (Ревзин 1977), составляющих текст, изучение образцов, по которым они формируются, рассмотрение многообразных отношений между ними. Единицы текста (чаще всего их называют «сверхфразовые единства» – впервые употребляются в работах Л.А. Булаховского, «сложное синтаксическое целое» – в работах Н.С. Поспелова) представляют собой важное звено, связующее план содержания и план выражения. Гипотеза о существовании и структуре этой единицы, ее связности получила дальнейшее развитие в трудах И.А. Фигуровского, В.В Виноградова, И.Р. Гальперина, Г.Я. Солганика и многих других.

§ 1.2. Текст и дискурс

Начиная с 60-х – 70-х годов прошлого века лингвистические исследования все чаще стали учитывать экстралингвистические данные, в частности, сведения о соотношении языка и познавательных процессов, особенностях восприятия и переработки информации, роли «фоновых» знаний в процессе понимания текста и т.д. Лингвистические исследования приобрели прагматический характер, к анализу стал привлекаться социальный контекст, появились термины «дискурс» и «дискурс-анализ».

Для лингвистики текста существенным является вопрос о тексте как процессе,[8] и здесь структурная модель описания текста как

[8] См. об этом подробнее: Карасик В.Н. Языковой круг: личность, концепты, дискурс. – М.: Гнозис, 2004 – С. 226–227.

самодостаточного герметичного образования становится недостаточной, возникает необходимость учета обстоятельств общения и характеристик коммуникантов, то есть требуется переход к коммуникативной модели представления текста. Такой переход, в частности, осуществляется в следующих направлениях:

1) осваиваются результаты исследований, так или иначе связанных с целым текстом, в прагма– , психо– и социолингвистике, риторике, литературоведении, когнитологии;
2) концептуально и терминологически противопоставляются текст, погруженный в ситуацию реального общения, то есть дискурс, и текст вне такой ситуации;
3) на первый план выходят вопросы, связанные с порождением и пониманием текста, с диалогической природой общения;
4) исследуются не идеальные, правильно построенные тексты, а текстовые стратегии в их разнообразных реализациях.

Еще более широкий круг вопросов, затрагивающих сущность текста как феномена человеческой культуры, рассматривается при лингвокультурологическом исследовании текста. Корни такого подхода прослеживаются в трудах В. фон Гумбольдта и его последователей, в том числе представителей лингвистической антропологии. Такой подход направлен на освещение особенностей менталитета народа, обусловленных его историей и отраженных в языке, прецедентных текстах, концептосфере и культурных концептах.[9]

Автор термина «дискурс» – Ю.Хабермас (Habermas 1981) – использовал его для обозначения речевой коммуникации, предполагающей рациональное критическое рассмотрение норм, ценностей и правил социальной жизни. Соотношение понятий «дискурс», «текст» и «речь» дискутируется давно и с неизменным интересом. Иногда их разграничивают по оппозиции письменный текст vs устный дискурс (Макаров 2003), что неоправданно сужает объем данных терминов, сводя их к двум формам языковой действительности – использующей и не использующей письмо. Такой

[9] Подробнее о концептах и концептосфере см. Лихачев 1993 и Степанов 1997.

подход весьма характерен для ряда формальных подходов к исследованию языка и речи. На этом основании исследователи склонны разграничивать дискурс-анализ, объектом которого должна быть лишь устная речь, и лингвистику (письменного текста). Но данное ограничение не срабатывает во многих случаях.

Изучению дискурса посвящено множество исследований, авторы которых трактуют это явление в столь различных научных системах, что само по себе понятие «дискурс» стало шире понятия «язык». М. Стаббс (Stubbs 1989) выделяет три основных характеристики дискурса:

1) в формальном отношении – это единицы языка, превосходящие по объему предложение;
2) в содержательном плане дискурс связан с использованием языка в социальном контексте;
3) по своей организации дискурс интерактивен, то есть диалогичен.

Дебора Шиффрин (Schiffrin 1994) выделяет три подхода к определению этого явления:

1) произведение, превышающее по объему предложение (формальная лингвистика);
2) любое употребление языка (функциональная точка зрения);
3) совокупность формально организованных контекстуальных единиц употребления языка.

В.Г. Костомаров и Н.Д. Бурвикова (1999) противопоставляют дискурсию (процесс развертывания текста в сознании получателей информации) и дискурс (результат восприятия текста, когда воспринимаемый смысл совпадает с замыслом отправителя текста). Такое понимание соответствует логико-философской традиции, согласно которой противопоставляются дискурсивное и интуитивное знания, то есть знания, полученные в результате рассуждения и в результате озарения.

Приоритет в исследовании социального контекста в описании дискурса принадлежит датскому лингвисту Т.А. ван Дейку, утверждающему, что дискурс – это «существенная составляющая социокультурного взаимодействия, характерные черты которого – интересы, цели и стили». Изменения и ограничения находят свое

проявление в дискурсе в виде определенных тематических репертуаров. Это значит, что пользователи языка могут формировать гипотезы относительно того, что будет или может быть сказано, кем и в какой ситуации.

В статье «Эпизодические модели в обработке дискурса» (ван Дейк 1997) Т.А. ван Дейк отмечает, что выражения естественного языка вообще и дискурс в частности могут употребляться для того, чтобы указывать на что-либо, обозначать что-либо «в мире» или в некотором социокультурном контексте. Дискурс «дает представление о предметах или людях, об их свойствах и отношениях, о событиях или действиях или об их сложном сплетении, то есть о некотором фрагменте мира, который он именует ситуацией».

Общий социальный контекст (*general social context*) ученый описывает в следующих категориях:

- личное;
- общественное;
- институциональное / формальное;
- неформальное.

Данные категории характеризуют различные виды социальных контекстов, например, общественные институты (суды, больницы и т.п.), неформальные общественные «места» (рестораны и т.п.), частные институты (семья), неформальные личные ситуации (драки, объяснения в любви).

Социальные контексты могут быть подвергнуты дальнейшему анализу в терминах следующих категорий:

- позиции (роли, статусы и т.п.);
- свойства (пол, возраст и т.п.);
- отношения (превосходство, авторитет и т.п.);
- функции (отец, продавец, судья и т.д.).

Характеристики социальных контекстов и характеристики их участников связаны между собой определенным образом. Они задают возможные действия участников социального взаимодействия в тех или иных ситуациях.

Социальным контекстам может быть придана определенная организация, например, с помощью некоторой структуры (социальных) фреймов. Так, например, внутри институциональной

ситуации суда имеется несколько фреймов, которые хронологически упорядочены: фреймы обвинения, защиты, и вынесения приговора. В этих фреймах участникам приписываются специфические функции, позиции, качества и отношения. Фреймы определяют также, какие виды действия могут быть совершены. Функции задают набор возможных социальных действий. Таким образом, чтобы описать социальные контексты, необходим набор конвенциальных установлений (*conventions* – правил, законов, принципов, норм, ценностей), которые бы определяли, какие действия ассоциируются с конкретными позициями, функциями и т.д.

Французская лингвистическая школа связывает понятие дискурса с исследованиями Э. Бенвениста, который называет дискурсом речь, присваиваемую говорящим, в противоположность повествованию, которое разворачивается без эксплицитного вмешательства субъекта повествования. Составляющими дискурса Э. Бенвенист называет отдельные и каждый раз единственные в своем роде акты, которыми говорящий активизирует язык в речь.

Сопоставляя различные подходы к пониманию дискурса, М.Л. Макаров (1998) намечает основные координаты, с помощью которых определяется дискурс: формальная, функциональная и ситуативная интерпретации. Формальная интерпретация – понимание дискурса как образования выше уровня предложения. Речь идет о сверхфразовом единстве, сложном синтаксическом целом, выражаемом как абзац или кортеж реплик в диалоге – на первый план здесь выдвигается система коннекторов, обеспечивающая целостность этого образования. Функциональная интерпретация в самом широком понимании – это понимание дискурса как использования языка, то есть речи во всех ее разновидностях.

В.Е. Чернявская (2001), рассматривая различные трактовки дискурса в отечественном и зарубежном языкознании, выделяет два его основных типа:[10]

[10] «Тип дискурса – это обобщенное представление о тексте, концепт текста в сознании носителей соответствующей культуры. В этой связи представляется обоснованным выделение в структуре дискурса трех компонентов: обобщенная модель референтной ситуации; репрезентация знаний о социальном контексте, с учетом которого осуществляется социальное взаимодействие посредством текстов; лингвистические знания». – Левицкий Ю.А. Проблема типологии текстов. – Пермь: Изд-во Перм. ун-

- конкретное коммуникативное событие, фиксируемое в письменных текстах и устной речи, осуществляемое в определенном когнитивно и типологически обусловленном коммуникативном пространстве;
- совокупность тематически отнесенных текстов.

Известное определение Н.Д. Арутюновой дискурса – это речь, «погруженная в жизнь» – предполагает, что термин «дискурс» не применим к древним и иным текстам, связи которых с живой жизнью не восстанавливаются непосредственно.[11] Дискурс является центральным моментом человеческой жизни «в языке», то есть языковым существованием: «Всякий акт употребления языка – будь это произведение высокой ценности или мимолетная реплика в диалоге – представляет собой частицу непрерывно движущегося потока человеческого опыта. В этом качестве он вбирает в себя и отражает в себе уникальное стечение обстоятельств, при которых и для которых он был создан».[12] К этим обстоятельствам относятся:

- коммуникативные намерения автора;
- взаимоотношения авторов и адресатов;
- всевозможные «обстоятельства», значимые и случайные;
- общие идеологические черты и стилистический климат эпохи в целом и той конкретной среды и конкретных личностей, которым сообщение прямо или косвенно адресовано, в частности;
- жанровые и стилистические черты как самого общения, так и той коммуникативной ситуации, в которой оно включается;
- множество ассоциаций с предыдущим опытом, так или иначе попавших в орбиту данного языкового действия.

В работе «Языковой круг: личность, концепты, дискурс» В.Н. Карасик (2004) выделяет и обосновывает категории дискурса с

та, 1998. – С. 85; Карасик В.Н. Языковой круг: Личность, концепты, дискурс. – М.: Гнозис, 2004. – С. 229–230.

[11] Арутюнова Н.Д. Дискурс // Лингвистический энциклопедический словарь. – М.: Сов. энцикл., 1990. – С. 136–137.

[12] Гаспаров Б.М. Язык, память, образ. Лингвистика языкового существования. – М.: Нов. лит. обозрение, 1996. – С. 10.

позиции коммуникативного языкознания, с учетом достижений как структурно-функциональной, так и культурологической лингвистики. Коммуникативный подход базируется на анализе коммуникативных обстоятельств как важнейшего смыслообразующего компонента текста. Автор выделяет следующие категории: участки общения (статусно-ролевые и ситуативно-коммуникативные характеристики); условия общения (пресуппозиции, сферы общения, хронотоп, коммуникативная среда); организация общения (мотивы, цели, стратегии, контроль общения и вариативность коммуникативных средств); способы общения (канал и режим, тональность, стиль и жанр общения).

Минимальной единицей коммуникативного взаимодействия следует считать двухстороннюю единицу: обмен, интерактивный блок, простая интеракция, элементарный цикл. В рассматриваемой выше работе Л.М. Макаров отдает предпочтение термину «обмен». Структурно обмены подразделяются на элементарные или простые (двухкомпонентные, двухшаговые обмены типа вопрос – ответ, просьба – обещание, приветствие – приветствие), и сложные или комплексные (типовые структуры, объединяющие три, четыре реплики, например, вопрос – ответ – подтверждение или вопрос – переспрос – уточняющий вопрос – ответ).

Более проблематичным выглядит выделение единиц дискурс-анализа, превышающих по объему сложные обмены. Некоторые авторы считают, что обмен – это уже высшая единица языкового общения. Тем не менее, оговаривая особый характер более крупного сегмента общения, многие исследователи ощущают необходимость выделения такой единицы. Одни авторы называют ее «трансакцией», другие – «фазой». Самым масштабным, и во многих случаях довольно легко идентифицируемым, структурным сегментом языкового общения, единицей макроуровня дискурса является то, что в этнографически ориентированной лингвистике трактуется как «речевое событие», в лингвистической прагматике – «макродиалог» или «макротекст». Примерами такой единицы могут быть урок в школе, заседание суда, деловое совещание, беседа и т.п.

Анализ дискурса предполагает его исследования с точки зрения когезии и когеренции. Когезия,[13] или формально-грамматическая связность дискурса, определяется различными типами языковых отношений между предложениями, составляющими текст или высказываемыми в дискурсе. М.А.К. Хэллидей и Р. Хасан[14] предлагают рассматривать пять аспектов таких отношений: указательную, личную и сравнительную референцию; субституцию имени, глагола и предикативной группы; эллипсис имени, глагола и предикативной группы; союзные слова и другие коннекторы, выражающие одно из ограниченного набора отношений, причем весьма общих, связывающие разные части текста; а так же лексическую когезию, часто достигаемую повтором одного и того же слова или лексического эквивалента исходного слова, повтора родового понятия, коллокации и так далее. К явлениям этого уровня относятся механизмы ко- и кросс-референции, анафоры и прономинализации, широко изучающиеся в лингвистике текста. Когеренция шире когезии, она охватывает не только формально-грамматические аспекты, связи высказываний, но и семантико-прагматические аспекты смысловой и деятельностной связанности дискурса как локальной, так и глобальной.

§ 1.3. Понятие жанра в отечественной науке

Традиции изучения жанров были заложены еще в трудах Аристотеля, который впервые ввел деление на литературные жанры (в работе *De poetica*) и речевые жанры (*Rhetorica*). В работе *Rhetorica* на основе элементов речевого акта (говорящий, тема и слушающий) он выделяет три речевых жанра: юридический дискурс, совещательный дискурс и церемониальный дискурс.

Каждый функциональный стиль оформляется в совокупность жанров – исторически сложившихся типов литературного произведения (художественного, публицистического, научного и др.),

[13] См. об этом подробнее: Макаров М.Л. Указ. соч. – С.194–197.
[14] Halliday, M.A.K., Hasan, R. Cohesion in English. – London: Longman, 2001. – P. 185-188.

например, роман, монография, репортаж и т.п.[15] Традиционно под литературными жанрами понимаются главным образом типы художественных произведений,[16] хотя каждая эпоха обладает своей системой жанров в каждой из сфер духовной деятельности и общения.

К.Ф. Седов (2007) считает речевой жанр одним из ключевых понятий, структурирующих новую научную парадигму: «Универсальность рассматриваемой категории, ее инструментально-методологический потенциал позволяет говорить о жанровом пространстве повседневной коммуникации как о самопрезентации и самораскрытия индивидуальных особенностей идиостиля (речевого портрета) личности, ее коммуникативной компетенции, составляющей которой выступает жанровая компетенция».

При этом В.В. Дементьев (2015) отмечает, что до сих пор не даны ответы на наиболее принципиальные вопросы (например, какие механизмы позволяют носителю языка идентифицировать речевые жанры в тех случаях, когда ни конкретная языковая форма реплик, ни их последовательность не имеют ничего общего с теми, с которыми он уже сталкивался в своей речевой практике, – часто высказываемая исследователями идея «ключевых» слов, опорных реплик или типических интенций не может быть эффективно применена во многих случаях, поскольку известные заранее «ключевые» конструкции и речевые фигуры могут вообще не встретиться во вполне гладко протекающем речевом общении); методика речежанровых исследований по степени формализации несопоставима с традиционными лингвистическими моделями.

В современной отечественной жанристике можно выделить два основных исследовательских направления: литературные жанры и речевые жанры. Вопрос о соотношении речевых и литературных

---

[15] См. Стилистический энциклопедический словарь русского языка / под ред. М.Н. Кожиной. – М.: Флинта: Наука, 2003. – С. 56.

[16] Жанр – исторически сложившаяся, удостоверенная традицией и тем самым наследуемая совокупность определенных тем и мотивов, закрепленных за определенной художественной формой, связывающая их между собой узнаваемыми чувствами и мыслями. (Гаспаров М.Л. Избранные статьи. – М.: НЛО, 1995. – С. 46.)

жанров остается неразрешенным. Наиболее аргументированной представляется точка зрения, согласно которой понятие речевого жанра является родовым по отношению к понятию литературный жанр, обозначающему не все реально существующие жанры, но лишь те из них, которые исторически признаны таковыми.

Литературные жанры обладают комплексом устойчивых свойств, причем эти свойства могут характеризовать самые различные стороны произведения – его объем, тематику, систему рифм, активность вымысла. Среди литературных жанров есть такие, которые бытуют на протяжении всей истории духовной культуры (например, басня); другие же существуют лишь в определенные эпохи. Признаки жанра эволюционируют: нередко одни и те же жанровые названия в разные периоды литературного развития выражают разные понятия, так как соотнесены с различными типами текста (к примеру, понятие оды во времена Ломоносова и в первой трети XIX в.).

Жанр по своей структуре неоднороден. Н.Т. Рымарь и В.Н. Скобелев (1994) отмечают жесткую закономерность художественного сознания, которая действует как система устойчивости аналогичных проявлений. «Жанровая структура» – это проникновение организованности в органику, знак, показатель, их слияние. Это и есть «ядро жанра, которое сохраняется в течение многовековой жизни жанра под слоем всяческих новообразований».[17] Жанровое «ядро»[18] как первооснова жанра, его инвариант, его неизменная сущность – фундамент, исходная точка вместе с тем и система ориентиров художественного развития. Именно в этой устойчивости, в этой инвариантности – залог развития, предварительное условие возможных изменений жанра в целом. Именно жанровая структура является залогом устойчивости того или иного жанра, и она связана с родовыми основаниями бытия любой личности.

---

[17] Рымарь Н.Т., Скобелев В.Н. Теория автора и проблема художественной деятельности. – Воронеж, 1994. – С. 127.

[18] Н.М. Разинкина использует понятие жанровой «доминанты», которая определяется через набор языковых признаков. (Разинкина Н.М. Функциональная стилистика. – М.: Высшая школа, 2004. – С. 10–11).

В отличие от жанров литературы, речевые жанры (например, дискуссия, лекция, консультация, доклад, беседа и т.д.) обычно строятся на многообразном чередовании или смешении, взаимопроникновении элементов разговорного и книжного языка. Речевые жанры[19] следует считать одним из ключевых понятий, структурирующих новую парадигму. Универсальность рассматриваемой категории позволяет говорить о жанровом пространстве повседневной коммуникации.

Впервые проблему жанров речи поставил крупнейший философ-филолог Михаил Михайлович Бахтин. Речевой жанр Бахтин (1996) считал категорией, которая позволяет связывать социальную реальность с реальностью языковой. Жанры речи он называл «приводными ремнями в истории общества к истории языка».

Согласно М.М. Бахтину (1986), речевые жанры – это «относительно устойчивые, композиционные и стилистические типы высказывания», «типовые модели построения речевого целого», которые для говорящего «имеют нормативное значение, не создаются им, а даны ему. Если бы речевых жанров не существовало <…>, речевое общение было бы невозможно».

Говоря о первичных (простых) и вторичных (сложных) речевых жанрах, М.М. Бахтин (1996) отмечает, что различие между ними «чрезвычайно велико и принципиально, но именно поэтому природа высказывания должна быть раскрыта и определена путем анализа и того и другого вида».

Для обозначения жанровых форм, представляющих собой одноактные высказывание, К.Ф. Седов (2007) предлагает термин «субжанр» – минимальные единицы типологии речевых жанров, равные одному речевому акту и выступающие в виде тактик, основное предназначение которых менять сюжетные повороты в развитии интеракции. А речевые формы, которые объединяют в своем составе несколько жанров, получили название «гипержанров».

[19] См. об этом подробнее: Антология речевых жанров: повседневная коммуникация / под ред. проф. К.Ф. Седова. – М.: Лабиринт, 2007.

Разработка категории речевого жанра привела к созданию особого перспективного направления антрополингвистики – жанроведения (генристики, генологии) (Арутюнова 1998; Барнет 1985; Вежбицка 1997; Гайда 1999; Гольдин 1997; Дементьев 1997, 1999, 2006; Китайгородская, Розанова 1999; Салимовский 2002; Седов 2001; Федосюк 1997; Шмелева 1990, 1995 и мн. др.).

В России сформировалось несколько центров (Волгоград, Екатеринбург, Краснодар, Красноярск, Москва, Новосибирск, Омск, Орел, Пермь, Санкт-Петербург, Саратов, Тверь), где последовательно и весьма успешно осуществляются описание и систематизация речевых жанров, а также разработка специального метаязыка для их описания.

Следуя духу концепции М.М Бахтина, представители этого направления определяют речевые жанры как «вербально-знаковое оформление типических ситуаций социального взаимодействия людей» (Седов 2007: 8).

В качестве основы для коммуникативной дифференциации жанров в зависимости от сферы человеческой деятельности они используют следующие оппозиции:

- письменный / устный, где структура письменных жанров тяготеет к более жесткой монологической форме, а устные жанры допускают большую вариативность в использовании языковых средств;
- официальный / неофициальный, где официальные жанры имеют большую степень конвенциональности и стереотипичности по сравнению с жанрами неофициальной коммуникации;
- публичный / непубличный, где жанры публичного общения предполагают более высокую степень осознанности в употреблении языковых средств, чем жанры непубличного общения.

Разные подходы к классификации жанров привели к постановке вопроса о разработке номенклатуры жанров и даже создании энциклопедии речевых жанров (Шмелева 2007). Однако на пути к

созданию такой универсальной схемы стоит проблема, которую А. Вежбицка (2007: 80) связывает с культурной спецификой того или иного языка, с тем, что жанры, выделенные, например, польским языковым сознанием, отражают польский общественнокультурный мир и кодифицируются в польских лексических единицах. Это позволяет автору утверждать, что речевые жанры, выделенные данным языком, являются одним из лучших ключей к культуре данного общества.

Отдельного внимания заслуживает вопрос о методике изучения речевых жанров. В.В. Дементьев (2015) связывает эту задачу с построением модели, включающей в себя такие аспекты, как:

- речежанровое содержание;
- языковое содержание;
- композиционное оформление высказываний.

В целом виде модель включает в себя следующий ряд параметров:

- предмет речи, (что́ именно, с какой целью, в каком контексте, с какой оценкой сообщают друг другу собеседники);
- стиль (коммуникативная тональность, где особенно важна степень серьезности);
- лексика – наименования жанров и их компонентов;
- особенности синтаксической структуры;
- целеполагание;
- социальный (включая аксиологический) фактор;
- общая внешнекультурная и внутрикультурная парадигма;
- основные коммуникативные сферы, где роль речевых жанров для данного периода является особенно критической;
- отдельные источники материала.

В целом можно отметить, что, несмотря на отдельные упоминания жанров делового или институционального общения, основное внимание в отечественной жанристике уделяется художественным и бытовым жанрам.

§ 1.4. Понятие жанра в зарубежной науке

Понятие «жанр» (и связанное с ним понятие «жанровый анализ») представляет собой достаточно обширную и не всегда четко очерченную область исследования современного английского языка. Само слово, звучащее в английском языке с характерным французским «акцентом», изначально относилось к популярному в XVIII-XIX вв. виду небольших живописных полотен, изображавших сцены из повседневной жизни.

В современной зарубежной (прежде всего, англоязычной) науке можно выделить три основные школы жанрового анализа:

– Новая Риторическая школа[20] (США),

– Австралийская школа,[21]

– школа ESP (English for Special Purposes).

Английский язык для специальных целей (ESP)[22] прошел в своем развитии ряд этапов, связанных с эволюцией самой науки о языке. Существенное влияние на ESP оказала теория дискурсивного анализа, которая позже положила начало и жанровому анализу. Поворотным моментом стали публикации работ Джона Суэйлза (1981), в которых автором была предложена подробная теоретическая схема жанрового анализа на материале раздела введения к научным статьям.

Исследование жанра в ESP можно хронологически разделить на две основные стадии:

– ранние работы, основанные на анализе риторических ходов и шагов (*moves & steps*), задействованных в дискурсе, – направление которое можно охарактеризовать, как структурно-аналитическое (*structural move analysis*);

– более поздние работы, расширяющие понимание жанрового анализа за счет рассмотрения влияния экстралингвистических (а позже – межкультурных) факторов, определяющих выбор языковых средств и последовательность элементов (sequencing).

---

[20] См. в связи с этим: Кибрик, Плунгян 2002.

[21] В частности, о жанрах применительно к теории речевых актов см. статью А. Вежбицкой в *Антологии речевых жанров* (2007).

[22] Подробно о содержании и истории ESP см.: Hutchinson 1995.

Теория жанрового анализа, предложенная Джоном Суэйлзом[23] (Swales 1990, 2004), включает в себя ряд базовых постулатов в отношении понятия «жанр»:

1. Жанр включает в себя **класс коммуникативных событий** (*class of communicative events*), к которым относятся такие ситуации общения, где язык играет определяющую роль. Виды человеческой деятельности, где разговор является случайным (напр., физические упражнения, домашние дела, вождение автомобиля и т.п.), а также любая невербальная зрительная или слуховая активность (напр., просмотр картин, прослушивание музыки и пр.) к коммуникативным событиям не относятся. Коммуникативные события отличаются по своей частотности – от широко распространенных (обслуживание клиентов, новостные заметки в газетах и др.) до крайне редких (напр., энциклика папы римского или пресс-конференция президента страны). При этом, последние, для того чтобы выделяться как жанр, должны обладать особой значимостью в данной культуре – если событие происходит лишь раз в год, ему необходимо обратить на себя большое внимание.

2. Коммуникативные события должны обладать таким признаком, как **общий набор коммуникативных целей** (*shared set of communicative purposes*). Внимание к коммуникативным целям и выделение их в качестве главного идентифицирующего признака жанра Джон Суэйлз объясняет тем, что, за редким исключением, жанры представляют собой ни что иное как коммуникативное средство достижения поставленных задач. То, что эти цели сложно определить и вычленить, ставит перед исследователем новые задачи, решение которых требует независимости и широты взглядов, ограждая его от упрощенного подхода, ограничивающегося инвентаризацией языковых и стилистических факторов. Нередко один жанр может подразумевать целый *набор* коммуникативных целей – напр., радионовости предназначены, в первую очередь,

[23] Определение жанра Дж.Суэйлза звучит следующим образом: A genre comprises a class of communicative events, the members of which share some set of communicative purposes (1990:58).

информировать слушателей о последних событиях в мире, но кроме этого они ставят перед собой такие задачи, как формирование общественного мнения, организация действий населения (к примеру, во время чрезвычайных ситуаций) и т.п.

3. Образцы или примеры жанров отличаются с точки зрения **соответствия прототипу** (*prototypicality*). Для выбора критериев отнесения к тому или иному жанру традиционно используются два подхода: *дефиниционный* (хорошо известный и использующийся в словарях, энциклопедиях и разнообразных научных текстах) и базирующийся на *семейном сходстве* (где основанием для объединения объектов в одну «семью» является не общий набор определяющих признаков, а другие, более свободные типы отношений). Третий, *прототипический* (или эталонный) подход основан на том, что представители одного класса имеют разный статус по отношению к данному классу.[24] Разнообразные факторы, такие как форма, структура и ожидания аудитории, служат для выявления того, насколько образец соответствует прототипу данного жанра.

4. Ограничения на допустимую свободу автора с точки зрения содержания, композиции и формы жанра определяются **логическим обоснованием** (*rationale*), «подоплекой» последнего. Авторитетные члены дискурсивного сообщества используют жанры для реализации своих целей. Таким образом, общий набор целей какого-либо жанра 1) признается авторитетными членами исходного (*parent*) дискурсивного сообщества, что может осознаваться ими в той или иной степени; 2) может только частично осознаваться младшими членами данного сообщества; 3) может осознаваться или не осознаваться посторонними. Осознание целей дает логическое обоснование жанра, а оно – в свою очередь – устанавливает ограничения. Джон Суэйлз иллюстрирует данное положение

[24] Например, одни птицы – в «большей степени птицы», чем другие. Так в североамериканской культуре первое, что ассоциируется со словом «птица», это малиновка, а не, к примеру, страус. У малиновки средний размер и традиционная форма, она летает, порхает по деревьям и поет. Это позволяет отнести ее к прототипу, эталону птицы в культуре США.

примером двух жанров административной корреспонденции. Сама административная корреспонденция жанром не является, но представлена рядом индивидуальных жанров, среди которых можно выделить «письмо с хорошей новостью» (‘*good news’ letter*) и «письмо с плохой новостью» *(‘bad news’ letter*). Оба представляют собой формальный ответ, например, на заявку на получение стипендии, субсидии или должности. На первый взгляд их можно отнести к одному жанру – «ответ на заявку», но более внимательный анализ показывает, что, несмотря на то, что они входят в один регистр и одно текстовое окружение, их различное логическое обоснование заставляет выделить для каждого свой жанр.

Логическое обоснование письма с хорошей новостью исходит, в первую очередь, из того, что информация будет воспринята с радостью. Поэтому новость излагают сразу и с энтузиазмом, в то время как остальная часть письма построена так, чтобы устранить все оставшиеся препятствия, ускорить и облегчить обратную связь. Отчасти логическим обоснованием такого письма является то, что коммуникация будет продолжена. При составлении письма с плохой новостью, наоборот, исходят из того, что информация не обрадует читателя. Поэтому саму новость помещают после некоего «буфера», который готовит получателя к разочарованию и состоит из выражений сожаления, не давая при этом оценок. Одной частью логического обоснования такого письма является минимизация чувства обиды, чтобы не осталось длительных отрицательных эмоций, связанных с этой организацией. Другая часть сигнализирует о том, что коммуникация окончена. С этой целью создаеся впечатление того, что отрицательное решение было принято не конкретным человеком, а неким обезличенным «комитетом», на который автор письма имеет мало влияния. Это делается для того, чтобы предотвратить возможные апелляции, жалобы и т.д. Таким образом, логическое обоснование определяет схематическую структуру дискурса и накладывает ограничения на выбор лексических и синтаксических средств.

5. **Номенклатура** (*nomenclature*) жанров данного дискурсивного сообщества является важным инструментом проникновения в их суть.

Знание общепринятых норм в отношении того или иного жанра (а также его логического обоснования) гораздо глубже у тех, кто постоянно или профессионально действует в рамках этого жанра, по сравнению с теми, кто сталкивается с ним случайно, время от времени. Вследствие этого активные члены дискурсивного сообщества становятся главными экспертами в области данного жанра, что можно наблюдать во время общения между профессионалами и их клиентами. Это приводит к тому, что они дают жанровые названия классам коммуникативных событий, которые, по их мнению, представляют собой воспроизводимые риторические действия. Эти названия зачастую заимствуются смежными или близкими сообществами, а затем – и более далекими и широкими.

Последователь Дж. Суэйлза, Тони Дадли-Эванс, подчеркивает в своих работах (Dudley-Evans, St John 1998) необходимость разработки системы анализа, которая бы наглядно показала отличие одного жанра от другого.[25]

Важным вкладом Дж. Суэйлза в жанровый анализ стала разработанная им методика выявления когнитивной структуры жанра на основе выделения в текстах одного жанра общего набора **риторических ходов** (*moves*)[26]. Ход – это часть текста, обусловленная или ограниченная особой коммуникативной функцией, т.е. такая часть текста, которая служит для достижения конкретной цели. Если

[25] Эта система, по мнению автора, должна включать в себя следующие составляющие: 1) объединение текстов, схожих между собой по риторической цели, форме и аудитории; 2) выявление того, чем данные тексты отличаются от других текстов и друг от друга; 3) сведения о риторической структуре и форме текста, которые могут использоваться при обучении языку.

[26] В отечественной литературе по анализу дискурса в качестве эквивалента англоязычного *move* используется термин «коммуникативный ход» или «коммуникативный акт» (т.е. речевой акт как единица диалогового взаимодействия в динамической модели речевой коммуникации – См.: *Англо-русский словарь по лингвистике и семиотике*. Представляется, однако, что данное понятие, главным образом, принадлежит сфере конверсационного анализа как «вербальное или невербальное действие одного из участников коммуникативного акта, минимальный значимый элемент, *развивающий взаимодействие*, продвигающий общение к достижению общей коммуникативной цели» (Макаров 2003:183), чему в английском языке соответствует термин *turn*.

каждый жанр обладает набором коммуникативных целей, то ход представляет собой риторический инструмент, реализующей фрагмент такого набора. Ход может состоять из единиц более низкого уровня – **риторических шагов** (*steps*).[27] В частности, в уже упоминавшейся работе 1981 года на основе анализа 48 текстов раздела «Введение» научных статей из абсолютно разных областей знания (от физики и биологии до социологии и лингвистики) автор обнаруживает у них общую структуру, состоящую из четырех шагов:

Шаг 1: **Определение поля исследования** (*Establishing the research field*);

Шаг 2: **Обобщение предыдущих исследований** (*Summarizing previous research*);

Шаг 3: **Подведение к новому исследованию** (*Preparing for the present research*);

Шаг 4: **Ознакомление читателя с новым исследованием** (*Introducing the present research*).

Учеником и последователем Джона Суэйлза является Виджай Бхатиа, исследующий не только жанры научной речи, но и другие профессиональные языки, в том числе юридический и деловой. Интересным представляется его исследование писем коммерческой рекламы (*Sales promotion letters*) с точки зрения их коммуникативной цели и когнитивной структуры (Bhatia 1993). Под указанным типом текста автор понимает незатребованное, написанное по собственной инициативе, письмо, адресованное избранной группе потенциальных клиентов (которые могут быть частными лицами или компаниями) с целью убедить их приобрести товар или услугу. Типичное рекламное письмо призвано служить достижению следующих коммуникативных целей:

1. Главная функция такого письма – персуазивная, убеждающая, в том смысле, что автор рассчитывает получить ответ. Поскольку в

[27] Moves and steps are “rhetorical instruments that realize a sub-set of specific communicative purposes associated with a genre” (Bhatia 2001:84).

данной ситуации эта задача является сложной, для обеспечения прагматического успеха необходимы следующие дополнительные коммуникативные цели:

2. Письмо должно привлечь внимание получателя, даже если на данный момент у него нет интереса к предлагаемому товару или услуге.

3. Рекламные письма обычно адресованы тем потенциальным клиентам, у которых – по информации автора – должна быть необходимость (сейчас или в будущем) в рекламируемом товаре или услуге. Поэтому главная функция такого письма – расхвалить товар или услугу на предмет того, насколько полезными они будут для покупателя.

4. Поскольку, как уже отмечалось, большинство рекламных писем пишутся по инициативе автора, а не по запросу получателя, а занятые бизнесмены не любят тратить свое драгоценное время на чтение рекламы, такие письма должны быть краткими и эффективными. Это требование вступает в противоречие с другим – в письме должна содержаться достаточная информация о товаре или услуге для тех клиентов, у которых уже есть в них необходимость.

5. Рекламные письма должны служить первым связующим звеном между потенциальным продавцом и покупателем, тем самым инициируя деловые отношения между ними. Поэтому все рекламные письма должны стимулировать дальнейшее общение между двумя сторонами.

В. Бхатиа анализирует структуру коммерческого рекламного письма банка *Standard Bank* (см. ниже) с точки зрения того, каким образом используемые автором риторические шаги служат достижению коммуникативных целей.

STANDARD BANK
268 Orchard Road, Yen Sun Building, Singapore 0923

4 December 1987
Mr Albert Chan
1 Sophia Road, 05-06
Peace Centre
Singapore 0922

Dear Sir

We are expertly aware that international financial man–agers need to be able to ask the right questions and work in the market place with confidence.	**Establishing credentials**
Corporate Treasury Services, Standard Bank, now provides a week-long Treasury Training programme designed to develop awareness and confidence in man-agers.	**Introducing the offer** *Offering product / service*
We explain the mechanics of foreign exchange and money markets. We discuss risk from an overall stand-point and practical hedging techniques to manage for-eign exchange risks. We also discuss treasury management information systems, taxation and the lat-est treasury techniques. We will be holding our next Treasury Training Programme from 24-28 February 1987, inclusive. The fee for the Training Programme will be US$1,500 per person to include all luncheons and a dinner as indicated in the schedule as well as all course materials.	*Essential detailing of the offer*
The programme is both rigorous and flexible. It can be tailored to fit the needs of a whole corporation or just a few levels within the company.	*Indicating value of the offer*
We are pleased to inform you that if your company sponsors 6 or more staff for the course, we will offer you a discount of US$100 per person.	**Offering incentives**
For your convenience, I enclose a reservation form which should be completed and returned directly to me.	**Enclosing documents**
If you have any questions or would like to discuss the programme in more detail, please do not hesitate to con-tact me (Telephone No. 532 6488 / telex No. 29052).	**Soliciting response**
As the number of participants at each training programme is limited, we would urge you to finalize as soon as possible your plans to participate.	**Using pressure tactics**
Thank you very much for your kind consideration.	**Ending politely**

Yours faithfully
Mr. G. Huff

В результате выявляется следующая структура:

1. **Обоснование своих полномочий** (*Establishing credentials*);
2. **Ознакомление читателя с предложением** (*Introducing the offer*):
    а) предложение товара или услуги;
    б) необходимая подробная информация о предложении;
    в) указание на ценность предложения;
3. **Предложение поощрительных стимулов** (*Offering incentives*);
4. **Прилагаемые документы** (*Enclosing documents*);
5. **Просьба ответить** (*Soliciting response*);
6. **Использование тактик давления** (*Using pressure tactics*);
7. **Вежливое окончание** (*Ending politely*).

В начале письма автор указывает, что адресат нуждается в его услугах. Очевидно, что такие письма адресованы тем лишь компаниям, в которых имеются финансовые менеджеры и которые могут извлечь пользу от их обучения. Автор письма заявляет, что его фирма способна удовлетворить потребности клиента в таком обучении. Фраза "*We are expertly aware…*" дает читателю понять, что он может быть уверен в профессионализме адресанта. Этот первый шаг В. Бхатиа называет **Establishing credentials**.

Поскольку большинство читателей таких писем не имеют желания приобретать рекламируемый товар или услугу, перед автором стоит трудная задача не только привлечь внимание читателя, но и убедить его в достоинствах своего товара. Один из способов – создать впечатление о репутации компании, о ее устойчивом положении на рынке, достижениях, специализации, многолетней успешной работе в данной конкретной области. Для этого авторы прибегают к повествованию от 1-го лица мн числа. (‘*we’ orientation*). И, наоборот, если хотят подчеркнуть свое понимание потребностей и интересов потенциального клиента, используется 2-е лицо *(‘you’ orientation*).

Убедив читателя в своих полномочиях, автор переходит к предложению своего товара. В следующих четырех абзацах он

подробно рассказывает о наиболее важных аспектах предлагаемых услуг: из чего они состоят, какова их стоимость, как они могут быть полезны читателю. В бизнесе это называется детализация товара, В. Бхатиа называет это **Introducing the offer**. Этот этап очень важен, поскольку, если читатель не знаком с товаром, его нельзя будет продать, каким бы хорошим он ни был. Данный ход состоит из трех шагов. Первым будет само предложение товара, вторым – его подробная характеристика, третьим – описание его достоинств. Расхваливая свой товар или услугу, авторы нередко прибегают к восторженным эпитетам (*lexical boosts*).

Предоставив читателю полную информацию об услуге, в шестом абзаце автор письма делает попытку сделать свое предложение еще более привлекательным, предлагая читателю стимул в виде скидки в размере 100 долларов при направлении на обучение шести или более сотрудников. Этот шаг называется **Offering incentives**. Однако это требование носит не столько универсальный, сколько культурно-специфичный характер, поскольку в одних странах принято торговаться, тогда как в других – нет.

Как уже отмечалось, авторы рекламных писем вынуждены искать компромисс между требованием краткости письма и необходимостью подробного информирования читателя. Одним из выходов из подобной ситуации является использование приложений (**Enclosing documents**) в виде брошюр, проспектов, бланков заявок и т.п.

Каждое рекламное письмо рассматривается как попытка установить деловые отношения или укрепить уже существующие. Поэтому их главной коммуникативной целью будет побудить читателя к продолжению общения. Для этого в письме указываются номера телефонов или имена людей, которые готовы ответить на все возникающие у читателя вопросы. Данный шаг (**Soliciting response**) характеризуется широким использованием формул вежливости.

Для того чтобы склонить еще колеблющегося читателя в свою сторону, авторы прибегают разнообразным тактикам давления – **Using pressure tactics**. К примеру, это может быть обещание скидок или других льгот при заказе товара ранее определенной даты. Отличие от

предыдущего шага (предложение стимулов) состоит в том, что если основная функция первого – убедить читателя в привлекательности товара или услуги, то у второго – подтолкнуть колеблющегося или сомневающегося клиента принять решение немедленно. Вот почему этот ход чаще всего встречается в конце письма.

В заключительной части рекламного письма необходимо вежливое окончание – **Ending politely**. Можно предположить, что любое деловое письмо должно оканчиваться вежливо, однако, степень вежливости может варьироваться в зависимости от жанра. Выделяют два типа дискурсивных функций окончания в деловых письмах: ситуационное (*situational*) и реляционное (*relational*). Ситуационные непосредственно связаны с функцией письма, тогда как реляционные выражают отношение автора письма к:

– будущему деловому сотрудничеству;

– будущим контактам;

– читателю письма.

Некоторые жанры (напр., научные) носят достаточно универсальный характер, тогда как другие (в частности, деловые) демонстрируют заметные культурные отличия.[28] Анализ когнитивной структуры является лишь начальным этапом комплексного исследования жанра, включающего в себя разные уровни – от собственного языкового до экстралингвистического, интертекстуального, социального, культурного и кросс-культурного.

[28] Например, при анализе рекламных писем из двух разных источников – транснациональных компаний (преимущественно западных) и местных (в данном случае – сингапурских) – выяснилось, что в качестве обязательных использовались шаги 2 и 4. Что касается шага 1, то местные фирмы использовали его крайне редко. Это объясняется тем фактом, что в Сингапуре мало компаний с давней репутацией на рынке, славящихся своей надежностью и авторитетом. При этом местные компании уделяли больше внимания шагу 3, полагая, что дополнительные стимулы смогут больше заинтересовать потенциальных клиентов.

2. Жанровый анализ (на примере англоязычного делового дискурса)

§ 2.1. О конвенциях применительно к речевым жанрам

Современные исследования жанра исходят из идей М.М. Бахтина (1986), рассматривающего жанры как точку приложения двух сил – центробежной (дифференциация) и центростремительной (конвенция).

Термин «конвенция» (*англ*. convention) восходит к латинскому существительному *convention* (соглашение, договор), которое, в свою очередь, образовано от глагола *convenire* (соглашаться, договариваться). В английский язык слово пришло в XV в. из старофранцузского, а в русский – в начале XVII в. из польского. Языковая конвенция тесно связана с семиотической функцией языка, в которой форма языкового знака не связана с обозначаемой вещью или явлением непосредственно. Связь эта установлена произвольно, условно, конвенционально. В естественных языках конвенции устанавливались исторически, независимо от воли отдельных людей. В искусственных языках конвенция установлена так же искусственно. Промежуточное положение занимают профессиональные языки и язык науки. В отношении указанных типов текстов принято считать, что ограничения в плане выбора лексико-грамматических средств накладываются их регистровой принадлежностью (т.е. ситуационно обусловленной функциональной разновидностью языка в отличие от, например, диалекта, обусловленного территориально, или стиля, относящегося преимущественного к сфере художественной литературы), тогда как жанровые ограничения действуют на уровне дискурсивной структуры текста.

Тем не менее, входя в тот или иной регистр, жанры характеризуются собственным инвентарем лексико-грамматических средств. Кроме того, будучи явлением не только языковым (текстовым), но и социокультурным, тот или иной жанр характеризуется конвенциями другого (контекстуального) уровня.

Взаимоотношения между дискурсивными и профессиональными факторами представлены в схеме, предложенной Виджаем Бхатиа (2007):

Рис. 1. Реализация дискурса в профессиональном контексте

Жанровые конвенции предполагают наличие явных или неявных договоренностей, которые существуют в сознании отправителей и получателей текстов. В частности, Виктор Борисович Шкловский (1967) считал, что сам жанр – «конвенция, соглашение о значении и согласовании сигналов. Система должна быть ясна и автору и читателю».

Анализ жанровых конвенций представляет собой дедуктивный когнитивный процесс, включающий в себя несколько последовательных этапов:

1. Предварительный анализ. На основе личного опыта и когнитивных знаний жанровых конвенций читатель относит текст к тому или иному жанру.

2. Изучение литературы и материалов, относящихся к данному жанру, его особенностям.

3. Анализ контекста. Данный этап включает в себя выявление отношений между адресантом и адресатом, ситуационные и межкультурные аспекты, напр., национальная или организационная культура, параметры коммуникации: средство, канал, время, место и т.п.

4. Собственно жанровый анализ. Выясняются критерии определения жанра (субжанров) рассматриваемого текста. Данный процесс проводится по трем уровням:

Уровень 1 – коммуникативная цель. Коммуникативная цель является ключевым детерминантом и играет главную роль в структурировании жанра. Цель выявляется на основе анализа коммуникативных ходов текста.

Уровень 2 – структура коммуникативных ходов. Текст как социальный процесс исходит из четкой цели и движется по определенным этапам, для которых авторы используют соответствующие коммуникативные ходы, одни из которых являются обязательными, а другие – факультативными.

Например, анализ такого жанра делового письма, как *letter of application* (заявление с просьбой о рассмотрении отправителя письма в качестве кандидатуры на вакантную должность), выявил следующие общие закономерности в структуре его основной части:

1. Обоснование своих полномочий (*Establishing credentials*).

2. Ознакомление читателя с кандидатурой соискателя (*Introducing the candidate*):

– предложение кандидатуры соискателя (*Offering the candidate*);

– необходимая подробная информация о соискателе (*Essential detailing of the candidate*);

– указание на ценность соискателя *(Indicating value of the candidate*).

3. Предложение поощрительных стимулов (*Offering incentives*).

4. Прилагаемые документы (*Enclosed documents*).

5. Использование тактики «давления» (*Using pressure tactics*).

6. Вежливое окончание (*Ending politely*)**.**

7. Просьба ответить (*Soliciting response*).

Уровень 3 – риторические стратегии. Анализ используемых в тексте риторических и визуальных стратегий, напр., частотность тех или иных синтаксических конструкций, лексических единиц или иконических знаков, подтверждают и усиливают первоначальное интуитивное восприятие текста (Толстова 2009). Ниже перечислены наиболее распространенные риторические стратегии:

– словарный состав (профессионализмы, терминология и т.д.);

– коннотация;

– переключение кода (диалект, социолект);

– образная система (метафоры, эпитеты и пр.);

– специфические синтаксические структуры (эллипсис, повтор, параллелизм и т.д.);

– специфические грамматические формы (номинализация, страдательный залог, действительный залог);

– специфические типы повествования (описание, нарратив, объяснение, аргументация и т.п.);

– интертекстуальность (апелляция к знаниям других текстов или фрагментов текстов);

– интердискурсивность (включение или имитация других жанров).

Узнаваемость является определяющим признаком конвенциональности. Само название жанра вызывает у адресата соответствующие ожидания. Оправдаются они или нет – зависит от присутствия в тексте специфического набора жанровых конвенций, который создает очертания жанра, делая его узнаваемым как для членов соответствующего коммуникативного сообщества, так и для непрофессионалов. Данный набор может количественно варьироваться – от наиболее полного (прототипического) до минимального (фрейма).

§ 2.2. Культурные особенности жанров делового английского

Не будет преувеличением утверждать, что взаимодействие языка и культуры проявляется на всех уровнях языка и во всех ситуациях человеческой коммуникации. Люди, даже говоря на родном языке, не всегда чувствуют себя уверенно в определенных ситуациях общения, как устного, так и письменного. Обычно это вызвано недостаточным владением тем или иным речевым жанром, например, презентация, светская беседа, телефонный разговор, обращение с просьбой и т.п.[29]

При общении на неродном языке ситуация еще больше усугубляется. Знание лексики и грамматики того или иного языка оказывается недостаточным для свободного общения на нем. Более того, возникает интерференция родного языка, когда говорящий/пишущий пытается передать устойчивые формы высказывания, принятые в родной культуре, средствами другого языка. Иногда эффект получается комичным (Девкин 1998, Гудков 2003), но всегда жанровые ошибки препятствуют адекватному пониманию и достижению желаемой коммуникативной цели.

Автор известной методики сопоставления бизнес-культур Гирт Хофштеде (Hofstede 2001) отмечал наивность представлений о том, что менеджмент во всем мире одинаков или унифцируется под влиянием глобализации. Это справедливо не только по отношению к практике ведения бизнеса и деловому поведению, но и к различным бизнес-жанрам. Иными словами, владение разнообразными жанрами, как на родном, так и на иностранном языке, является важнейшим условием языковой и коммуникативной компетенции.

---

[29] М.М. Бахтин отмечал, что «речевая воля говорящего осуществляется прежде всего в выборе определенного речевого жанра. Этот выбор определяется спецификой данной сферы речевого общения, предметно-смысловыми (тематическими) соображениями, конкретной ситуацией речевого общения, персональным составом его участников и т.п. И дальше речевой замысел говорящего со всей его индивидуальностью и субъективностью применяется и приспособляется к избранному жанру, складывается и развивается в определенной жанровой форме» (Бахтин М.М. Проблемы речевых жанров // Литературно-критические статьи. – М.: Художественная литература, 1986. – С. 441-452).

При исследовании жанра следует четко разграничивать собственно текстовый анализ, связанный с лексикой, грамматикой, синтаксической когезией и структурой текста, и дискурсивный анализ, включающий в себя речевые акты, текстовую когеренцию и, наконец, интертекстуальность.

Ванда Орликовски и Джоанна Йейтс (Orlikowski, Yates 1994) отмечают, что формальные признаки жанра должны также соответствовать воспроизводимым ситуациям, в которых они встречаются. С точки зрения «формы» жанра авторы выделяют три аспекта: структура, средства и язык – каждый из них относится к разным признакам реализации текстовой субстанции. Первый аспект – структура – связан с рубрикацией текста, например, такие приемы форматирования, как наличие перечней и списков в письменных текстах, или стандартный шаблон повестки дня совещания и т.п.

Термин «средства» обозначает технологию создания, передачи или хранения сообщения, например, факса, электронной почты, «бумажного» письма и т.п. Средства различаются в зависимости от жанра текста, который они передают – форма аттестации сотрудника, технический отчет и т.п. Йейтс и Орликовски, однако, допускают взаимосвязь между средствами и жанром и что некоторые жанры могут привычно ассоциироваться с конкретными средствами (напр., письмо, написанное на бумаге). Кроме того, воспроизводимые ситуации требуют использования определенных средств: напр., электронное сообщение, написанное в ответ на электронное сообщение.

Последний, третий аспект – это язык, который авторы определяют как «лингвистические характеристики»: стиль (формальный/неформальный), специализированная лексика, профессиональный жаргон.

Семиотико-интерпретационный (Bargiela-Chiappini 1999) подход рассматривает жанр как многостороннее явление, а именно:

– *интерактивное* – жанр является дискурсивным действием, осуществляемым автором и читателем на стыке языка и контекста;

– *реляционное* – жанр предписывает автору и читателю принять на себя определенные роли (социальные, институциональные, эмоциональные и т.д.), необходимые для интерпретации и вербализации текста;

– *риторическое* – структура, форма, функция и значение рассматриваются не как имманентные признаки дискурса, а как результат непрерывного процесса его производства и восприятия;

– *многозначное* – жанр дает простор различным интерпретациям и, как объект, может быть представлен разнообразными видами (текст, изображение, различные цвета, материалы и т.п.);

– *гетерогенное* – нет «чистых» жанров, а есть лишь конвенциональный набор жанровых форм;

– *интертекстуальное* – структура и значение дискурса формируются предшествующими (реальными или воображаемыми) дискурсами, которые уже имели место и в которые автор «включает» свой новый дискурс, создавая таким образом конкурентную перспективу;

– *историческое* – жанр помещен внутри некоего семиотического пространства тех значений и атрибуций, которые участники (выступающие в той или иной роли) создают в конкретный момент времени.

В данный перечень следовало бы включить и *межкультурный* компонент, поскольку аналогичные жанры в разных лингвокультурах демонстрируют значительные расхождения. Рассмотрим проявление межкультурной асимметрии на примере некоторых распространенных жанров англоязычной деловой переписки: письмо-просьба (*Letter of Request*) и письмо-обращение на работу (*Application Letter*).

Исследования А. Киркпатрика (Kirkpatrick 1991), посвященные письмам-просьбам, написанным на английском языке (жителями США) и на диалекте мандарин китайского языка, показывают, что последовательность изложения информации в таких письмах определяется нормами вежливости, принятыми в данной культуре. Из-за различного понимания вежливости американцы и китайцы не

только по-разному структурируют свои письма, но и по-разному формулируют саму просьбу.

Китайцы предпочитают размещать формулировку просьбы в конце письма, которое обычно имеет следующую структуру: *приветствие, преамбула, причины* и, наконец, сама *просьба*. Это объясняется тем, что просьба сама по себе не очень приятна для получателя письма и необходима преамбула (т.е. обмен любезностями, светская беседа, служащая для «сохранения лица» и «поддержания престижа»), которая в данном типе писем выполняет две функции – поднять настроение адресата и «подсластить пилюлю» просьбы.

Американцы придерживаются другой схемы: *обращение, причина, просьба или обращение, просьба, объяснение*. Такой быстрый переход к делу связан с тем, что, поскольку с их точки зрения сама просьба уже является обременительной для получателя, следует избавить его еще и от необходимости читать длинное сообщение на мониторе компьютера. Таким образом, для китайцев вежливость проявляется в смягчении просьбы через включение дополнительной информации, тогда как для американцев – в краткости сообщения.

Различия наблюдаются и в выборе языковых средств реализации стратегий просьбы, среди которых выделяют восемь основных (Bargiela-Chiappini 1999:130-131):

1. **Императив** (e.g. *Move* your car.).
2. **Перформатив** – когда иллокутивное намерение эксплицируется автором с помощью специальных перформативных глаголов (e.g. I'm *asking* you to move your car.).
3. **Дистанцированный перформатив** – перформативный глагол сопровождается различными средствами дистанцирования, такими как модальные глаголы или глаголы намерения (e.g. I'd like to *ask* you to move your car.).
4. **Директив с модальностью долженствования** – иллокутивное намерение выражает обязательство, налагаемое на адресата (e.g. You'll *have to* move your car later.).

5. **Директив с модальностью желания** – высказывание выражает желание автора, чтобы названное действие было выполнено (e.g. I *want* you to move your car.).
6. **Приглашение к действию** – иллокутивное намерение выражено с помощью конвенциональной формулы приглашения к действию (e.g. *How about* moving your car a little bit?).
7. **Включение предварительного этапа** – высказывание содержит указание на предварительное условие или целесообразность просьбы, особенно с точки зрения возможности, желательности, вероятности, выражаемое специализированными языковыми средствами (e.g. *Would it be possible* for you to move your car a little bit? I wonder *if you could move* your car a little bit.).
8. **Намек** – иллокутивное намерение не выводится из высказывания непосредственно, а выражается иносказательно, имплицитно (e.g. Somebody needs to pass the road.).

Первые четыре стратегии являются наиболее категоричными и, следовательно, менее вежливыми, тогда как последние четыре представляют собой косвенные речевые стратегии, когда высказывания звучат менее категорично и потому более вежливо[30].

Анализ писем показал, что американцы предпочитают имплицитные стратегии, в то время как китайцы, пишущие письма на английском языке, склонны к использованию эксплицитных стратегий. Таким образом, американцам свойственна прямая манера структурирования излагаемой информации, но, в то же время, предпочтение отдается более вежливым, опосредованным языковым средствам.

Виджай Бхатиа (Bhatia 1993) анализирует межкультурные различия, проявляющиеся в структуре такого жанра как письмо-обращение на работу (*Application Letter*), написанное носителями английского языка и жителями стран, где английский является вторым государственным языком – Индии, Пакистана, Шри-Ланки, Бангладеш и Сингапура.

[30] О косвенных речевых актах см. Клюев Е.В. Речевая коммуникация. – М., 1998. – С. 205 – 206.

Необходимо сразу разграничить *Application Letter* и *Cover Letter* (сопроводительное письмо к резюме). В первом у автора есть возможность, используя различные средства и коммуникативные стратегии, воздействовать на адресата, отступая, таким образом, от максимы качества и акцентируя максиму релевантности[31]. При этом наблюдается определенное сходство данного жанра писем с рекламными письмами, содержащими предложение товара или услуги (*Sales Promotion Letters*), с точки зрения их когнитивной структуры.

Среди основных стратегий, используемых в рассматриваемом типе писем, В. Бхатиа называет следующие:

1. Восхваление самого себя (*Self-glorification*) – необязательно подкрепленное фактами заявление о собственном превосходстве. Обычно это связано с огромным желанием получить хорошую работу.

2. Восхваление собеседника (*Adversary glorification*) связано с традицией восточного гостеприимства. Автор письма превозносит компанию или организацию, в которую обращается, а иногда и страну, если речь идет о работе за рубежом. Эта стратегия особенно усиливается, когда адресант ищет работу в богатых странах.

3. Самоуничижение (*Self-degradation*) – автор намеренно принижает финансовые возможности своей организации или даже страны. Данная стратегия служит для подчеркивания разницы между теперешней ситуацией соискателя и той, к которой он стремится. Поскольку получатель письма является единственным человеком, способным изменить положение вещей, то цель данного коммуникативного хода – вызвать сочувствие или жалость, что является эффективной персуазивной стратегией. Самоуничижение и самовосхваление (вместе и по отдельности) служат для достижения одной и той же цели. На первый взгляд это можно объяснить экономической зависимостью от развитых стран. Однако аналогичные данные обнаруживаются и в письмах, авторы которых проживают в таких промышленно развитых и процветающих странах, как Сингапур.

[31] О максимах Грайса см.: Грайс 1985; Клюев Е.В. 1998; Агапова 2004.

Таким образом, даже хорошо владея системой языка (его фонетическим, лексическим и грамматическим аспектами), инофоны склонны совершать речевые ошибки, поскольку действуют в соответствии с социокультурными нормами, определяющими использование их родного языка.

Межкультурные отличия никоим образом не означают, что одна культура лучше или «правильнее» другой. Однако, действуя в том или ином культурном пространстве и стремясь при этом достичь определенной цели и произвести определенное впечатление, нельзя не учитывать специфику жанра, принятую в данной культуре или группе культур.

§ 2.3. Наджанровые и межжанровые связи в англоязычном деловом дискурсе

В реальной человеческой коммуникации различные жанры, как правило, существуют не изолированно, а в тесном взаимодействии с другими жанрами, являющимися их «соседями» по определенной коммуникативной ситуации. Совместное функционирование нескольких жанров предполагает выработку определенного стереотипного сценария поведения в той или иной ситуации общения – стереотип целеполагания поведения, восприятия, понимания, общения.[32] Для описания классов таких стереотипных ситуаций, входящих в национальную культуру на правах ее элементов, используется понятие «лингвокультурный сценарий».

Под лингвокультурным сценарием понимается инвариант класса типовых ситуаций национальной культуры, существующих как объективно (в составе культуры), так и субъективно (в коллективном тезаурусе носителей данной культуры или, в лингвистическом аспекте, в фоновых знаниях носителей данного языка).[33]

[32] См. об этом подробнее: Маслова В.А. Лингвокультурология. – М.: Академия, 2001.
[33] См. подробнее: Савицкий 2006.

Примером такого лингвокультурного сценария может служить ситуация трудоустройства. В настоящее время международное сотрудничество в сфере бизнеса и экономики неуклонно расширяется. Многие российские специалисты выезжают за рубеж для работы по контракту, иностранные фирмы открывают свои филиалы в нашей стране, и сотрудниками этих филиалов становятся российские граждане.

Первым шагом при обращении по устройству на работу является написание **автобиографии**, или **резюме** (*résumé* или *CV* – в англоязычном варианте). Сегодня многие работодатели получают тысячи резюме от потенциальных работников, поэтому очень важно, чтобы резюме выделялось из остальных. Первое впечатление играет очень важную роль, соответственно необходимо иметь хорошее и правильно построенное резюме.

Структура и стиль написания резюме определяется культурой страны, где соискатель пытается найти работу, недостаточно просто перевести текст с одного языка на другой. Эффективное резюме должно соответствовать стандартам, принятым в той или иной стране.

Существует несколько типов резюме, самыми распространенными из которых являются хронологическое резюме, функциональное резюме и смешанное резюме.

Резюме не должно быть слишком объемным и обычно включает в себя следующие разделы:

- личные данные: ФИО, возраст, пол, национальность, дата и место рождения, семейное положение;
- домашний адрес, номер контактного телефона и адрес электронной почты;
- сведения об образовании, период обучения, специализация, название учебного заведения и его местонахождение (страна, город); информация о наградах, грамотах и т.п.;

– опыт работы в следующем порядке: период работы, название компании, название отдела, занимаемая должность, краткое описание служебных обязанностей на каждой работе;
– профессиональные навыки;
– исследовательские работы/диссертации (с кратким описанием);
– знание иностранных языков: степень владения устной и письменной формой иностранного языка, наличие сертификатов;
– членство в профессиональных организациях;
– интересы (в т.ч. поездки в страну работодателя);
– личностные характеристики.

Помимо фактических сведений о соискателе резюме содержит «Личный портрет» (*Personal profile*), в котором дается характеристика его профессиональных и личностных качеств.

Анализ текстов резюме из открытых интернет-источников с практическими рекомендациями по написанию резюме[34], показал, что они имеют четкую структуру и состоят из следующих основных разделов:

– общие сведения;
– образование;
– опыт работы;

[34] Основные ресурсы: Resume Resourse [Электронный документ] (http://www.resume-resource.com), проверено 01.09.18; English Club [Электронный документ] (http://www.englishclub.com), проверено 01.09.18; CV-Template [Электронный документ] (http://www.cv-template.cn), проверено 01.09.18; DocStoc [Электронный документ] (http://www.docstoc.com), проверено 01.09.18; College of Art and Science [Электронный документ] (http://*www.physics.sc.edu),* проверено 01.09.18; *«Открытый урок»* [Электронный документ] (http://*www.festival.1september.ru*), проверено 01.09.18; *CVTips* [Электронный документ] (http://*www.cvtips.com*), проверено 01.09.18.

– умения и навыки;
– интересы;
– указание искомой должности.

Каждый раздел обслуживается определенным набором лексических и грамматических единиц. Наиболее продуктивными грамматическими структурами являются: формы прошедшего времени, инфинитивные и именные конструкции. Лексика представлена преимущественно именными частями речи и содержит единицы с положительной эмоциональной семантикой.

Следующим пунктом рассматриваемого лингвокультурного сценария является написание **письма,** в котором соискатель выражает свой интерес к вакансии, предлагает свою кандидатуру и к которому он прилагает составленное резюме. В английском языке используется термин **Letter of application** – заявление с просьбой о рассмотрении отправителя письма в качестве кандидатуры на вакантную должность[35], которое обычно пишется в ответ на увиденное объявление либо для того, чтобы узнать о наличии вакантной должности.

Анализ образцов *Application letters*, опубликованных на различных сайтах сети Интернет, посвященных проблематике трудоустройства в иностранных компаниях, с использованием методики анализа когнитивной структуры, предложенной Дж. Суэйлзом и В. Бхатиа, выявил определенную структуру писем-обращений по трудоустройству. Ниже приведен образец анализа структуры и выявленные коммуникативные ходы в одном из писем рассматриваемого жанра.[36]

[35] «Letter of application is a letter written by someone asking for a job, usually in response to an advertisement» – Назарова Т.Б. Словарь общеупотребительной терминологии английского языка делового общения. – М.: АСТ: Астрель, 2002. – С. 68.
[36] Источник: Career Service. University of Minnesota Duluth. [Электронный документ]. — http:// careers.d.umn.edu/cs_handbook. Проверено 01.09.2018.

I am writing to inquire about the possibility of securing a Pharmaceutical Sales Representative position with Cornier, Inc. My career focus is to become employed in the pharmaceutical sales field	**Ознакомление читателя с кандидатурой соискателя**
I have enclosed a resume outlining my qualifications.	**Прилагаемые документы**
My background consists of a wide range of sales, prospecting, and customer service experiences. I have an excellent track record in sales and making successful transitions to companies with varied products. Fast I started at Paine Webber as an intern, answering questions and request of clients. I was a learner quickly hired as a Sales Associate when they discovered my ability to generate leads by explaining the specifics of the stock market to clients. Driven I started my first sales job when I was 16 and have held a job continuously throughout college. Reliable I have a perfect employment record having never missed a day of work.	*Необходимая подробная информация о соискателе*
I have always had an interest in pharmaceutical and health related industries. Working in customer-oriented positions has enhanced my communication skills and my ability to generate sales without using high pressure tactics.	*Указание на ценность соискателя*
I can create significant interest in products by explaining their benefits and then successfully closing the transaction. I am also near completion of a four-year degree, demonstrating that I have the dedication to and can learn your sales program very quickly.	**Предложение поощрительных стимулов**
I would appreciate the opportunity to visit with a company representative regarding any Pharmaceutical Sales Representative positions. My plans are to relocate to the southern California area after graduation and I would prefer a position in the area.	**Использование тактики «давления»**
I will call you during the week of October 20, to discuss the possibilities. To contact me before then, email кathyjillian@yhoo.com or call 218-396-0000.	**Просьба ответить**
Thank you for your time and consideration.	**Вежливое окончание**

Анализ других писем выявил аналогичную структуру основной части письма, включающую в себя коммуникативные ходы, отмеченные в работе В. Бхатии (Bhatia 1993).

Особенностью писем-обращений на работу является рамочная конструкция: кандидат указывает на вакантную должность и выражает свое желание работать в данной организации в начале и в конце письма, например:

I would like to be a technical writer or an editor for Animal Publications. <…>

The possibility of working for Animal Publications as an editorial assistant is personally very exciting.

Как уже отмечалось (Bhatia 1993), наблюдается определенное сходство между письмами-обращения на работу (*Application letters*) и рекламными письмами с предложением товара или услуг (*Sales Promotion letters*), что проявляется в выполнении одинаковых коммуникативных функций (убеждение приобрести товар или услугу или принять кандидата) и в наличии общих компонентов их жанровой структуры, представленных коммуникативными ходами.

Ключевым элементом культурного сценария «устройство на работу» является интервью (**job interview**)[37] или собеседование. Оно представляет собой устную разновидность речи и носит двусторонний характер.

Собеседование представляет собой сложный процесс отбора кандидатов в тех или иных целях. Основная задача собеседования при наборе штата – увидеть, сможет ли кандидат влиться в сложившийся коллектив и успешно в нем работать.

Как и другие жанры, жанр собеседования имеет своеобразную организацию и собственный набор языковых средств. В учебных пособиях по управлению персоналом приводятся различные способы проведения интервью. Оно может быть структурированным, имеющим четкий план и список вопросов, и свободным,

[37] Interview – интервью, беседа, собеседование, деловое свидание, встреча. An interview for a job – собеседование при приеме на работу. – Колесникова Н.Л. Деловое общение. Business communication. – М.: Флинта: Наука, 2005. – С. 137.

неструктурированным. Важно отметить, что в зависимости от цели проведения собеседования могут использоваться различные методы его ведения:[38]

- метод регламентированного собеседования – собеседование проводится по строгим правилам; оно лишь немногим отличается от анкетирования. Данный метод применяется для проверки пригодности кандидата. Работодателю необходимо заранее составить перечень вопросов с использованием терминов, применяемых в данной отрасли;
- метод целенаправленной беседы с взаимным обменом информацией (самый распространенный метод) позволяет убедиться в том, что кандидат готов работать, прилагая максимум усилий при существующих условиях. При этом на собеседовании необходимо наиболее подробно объяснить характер и условия работы. Проводящий собеседование должен полностью знать содержание работы и квалификации, личных качеств и образования, а также особые требования, связанные с выполнением данной работы;
- метод свободного собеседования дает возможность убедиться в умении кандидата самостоятельно мыслить и находить ответы на вопросы до того, как они заданы. Используя данный метод, важно помнить, что, если поставленные задачи собеседования определяют характер требуемой информации, то содержание получаемой информации определяется задаваемыми вопросами.

Анализ текстов собеседований при приеме на работу из аутентичных пособий по английскому языку для специальных деловых целей (Emerson, P. Business Grammar Builder (2002); Evans, D. Powerhouse: An Intermediate Business English Course (1999); Jones, L., Alexander, R. New International Business English (2001), показал, что жанровая структура интервью представлена тремя основными

[38] Куланов М.Н. Управление кадрами: в помощь начинающему руководителю. – М.: Дашков и Ко, 2005. – С. 84–86.

составляющими: **вступительная часть** (*opening*), **основная часть** (*body*) и **заключительная часть** (*close*).[39]

Во время вступительной части (обычно она длится 2–5 минут) интервьюер старается снять волнение у кандидата. Некоторые работодатели задают общие вопросы о его интересах, другие рассказывают о предлагаемой вакансии или компании.

Основная часть (10–15 минут) – это возможность кандидата рассказать о своей квалификации, умениях и навыках. Интервьюер пытается выявить сильные и слабые стороны потенциального работника и узнать информацию, не включенную в резюме.

В заключительной части (2–5 минут) интервьюер говорит, что ожидает кандидата после интервью.

Основная часть строится по принципу вопрос-ответ, в большинстве случаев вопросы касаются образования, личных качеств кандидата, предыдущего места работы, планов на будущее и т.п.

В начале большинства собеседований задаются вопросы о предыдущем месте работы (**Asking for the previous experience**). Для этого чаще всего используется форма The Past Indefinite Tense (*And did you manage to find a job easily after you left university*?).

Следующий тип вопросов посвящен навыкам и способностям кандидата (**Asking for the skills and abilities**). Через них интервьюер пытается определить, соответствует ли кандидат требованиям корпоративной культуры (*Would you say that you're a natural communicator*?). Для данного типа вопросов характерны модальные конструкции, смягчающие их категоричность и облекающие их в более вежливую форму.

Одним из частотных типов вопросов являются вопросы о причинах ухода с предыдущего места работы (**Asking for the reasons of leaving the previous work***): So why are you leaving?* Необходимость в данных вопросах определяется стремлением интервьюера узнать обстоятельства смены места работы. В них используются глаголы в

[39] Locker, K.O. Business and Administrative Communication. – Irwin McGraw-Hill, 1998. – P. 560.

форме The Past Indefinite Tense и The Present Continuous Tense (если смена работы только планируется).

С помощью вопросов относительно планов кандидата на будущее (**Asking for the plans for the future**) работодатель пытается определить перспективность кандидата и его умение креативно мыслить и строить планы на будущее. Наиболее частотной формой, обслуживающей данную функцию, является The Present Indefinite Tense: *Where do you see yourself in let's say in five years' time?*

Неотъемлемой частью собеседования при трудоустройстве являются вопросы, задаваемые кандидатом. Обычно интервьюер сам приглашает его к этому (**Asking for the questions**): *Is there anything else you'd like to ask me?*

Собеседование является финальным и решающим этапом процесса трудоустройства. По его результатам решается вопрос о приеме кандидата на работу или отказе в работе.

Таким образом, можно наблюдать, что разные жанры демонстрируют ярко выраженные связи друг с другом – письмо-обращение на работу дополняется автобиографией или резюме, а за ними следует собеседование при приеме на работу. Каждый из этих жанров имеет четкую структуру и реализуется в соответствии с принятой в данном культурном сообществе форме. Вместе все три жанра составляют лингвокультурный сценарий, знание которого необходимо для успешного участия в ситуации трудоустройства.

Подводя итоги, в наиболее упрощенном виде можно сделать вывод о том, что, несмотря на некоторое пересечение в плане изучаемого объекта, лингвистика текста тяготеет к исследованию законченных (преимущественно письменных) текстов, дискурс-анализ включает в себя широкий контекст и участников общения (как устного, так и письменного), тогда как жанровый анализ (особенно в зарубежной практике) сосредоточен на когнитивной структуре прототипического варианта произведения речи. Особое внимание при этом уделяется жанрам профессионального, делового, академического общения.

3. Корпусная лингвистика – направление или метод?

§ 3.1. История компьютерных подходов к изучению языка

Мысль о том, что достоверные данные о фонетической, морфологической, синтаксической и семантической структуре языка и речи могут быть получены только из достаточно большого массива текстов, была высказана в 1965 году российским ученым профессором Р.Г. Пиотровским в докладе «Статистическое исследование лексики и грамматики текста с помощью электронно-вычислительной машины», прочитанном в Московском государственном педагогическом институте иностранных языков[40].

Однако еще раньше, в 1961 году, в работе «О точных методах исследования языка» под редакцией О.С. Ахмановой можно прочитать, что современные лингвистические задачи требуют принципиально нового подхода к языку, разработки особых методов исследования и описания языков. Этот подход автор определяет следующим образом: «Надо научиться представлять грамматические, лексические, лексико-фразеологические и другие закономерности языка в таком виде, чтобы их можно было, непосредственно подавать на приборы». Иными словами, необходимо «расшифровать» те процессы, посредством которых осуществляется языковое общение. Математика с ее неисчерпаемыми возможностями должна дать основания для гораздо более глубокого проникновения в «механизм» языка и вполне строгого и логического, вполне «рационального» описания открываемых ею закономерностей (Ахманова 1961: 6).

Н.Б. Гвишиани (1997) объясняет необходимость появления корпусных исследований (в основном, применительно к английскому языку) причинами как лингвистического, так и конкретно-исторического характера:

– потребность в совершенствовании практики перевода и уровня владения иностранными языками;

[40] См. об этом: Зубов 1996.

– возрождение интереса к сопоставительным исследованиям в условиях интернационализации и интеграции общества в странах Европы;

– потребность сохранения единой языковой культуры в условиях распространения английского языка в разных частях света;

– необходимость обобщения многообразных проявлений английского языка в его различных вариантах, как диалектных, так и стилистических (регистровых);

– необходимость сохранить тождество хотя бы письменной формы английского языка в условиях, когда страны, официально признающие английский язык вторым государственным языком, стремятся к «лингвистической независимости».

Первый большой корпус данных по английскому языку начал составляться в Университетском колледже Лондона в 1960-е годы. Он назывался “Survey of English Usage”, а руководил им Рэндольф Кверк. Материал состоял из одного миллиона слов, представленных в 200 устных и письменных текстах, каждый из которых содержал 5000 слов. Тексты записывались в транскрипции на карточки, которые обрабатывались вручную. Позднее Йэн Свартвик из Лундского университета сделал электронную версию собранного материала, и она получила название “London-Lund Corpus of Spoken English”.

Собственно компьютерный корпус был впервые создан в 60-е годы в Брауновском университете (г. Провиденс, штат Род Айленд, США). Его авторами были Генри Кучера и У. Нельсон Френсис. Этот корпус также состоял из одного миллиона слов (пятисот 2000-словных прозаических печатных текстов) американского варианта английского языка. Тексты принадлежали пятнадцати наиболее массовым жанрам англоязычной печатной прозы США, вышедшей в свет в течение одного года. Авторы употребили слово «корпус» в значении «совокупность текстов, считающаяся представительной для данного языка, диалекта или другого подмножества языка, предназначенная для лингвистического анализа» (Френсис, 1982). Корпус сопровождался не только обширным описанием, но и большим

количеством материалов его первичной статистической обработки – частотный и алфавитно-частотный словарь, разнообразные статистические распределения.

В 1970-е годы возник британско-норвежский проект, в результате которого был собран корпус под названием “Lancaster-Oslo/Bergen Corpus of British English” (сокращенно – LOB). Задачей его было сравнить Брауновский корпус с текстами на британском варианте английского языка.

Следующий корпус был разработан исключительно в лексикографических целях в 1980-е годы под руководством Джона Синклера и получил название “Collins-Birmingham University International Language Database” (сокращенно *COBUILD*). В нем использовались новые технологии, такие как сканирование, для считывания большого числа печатных текстов в добавление к уже имеющемуся на электронных носителях значительному количеству материала. Это привело к знаменательному прорыву теперь уже в лексикографии. Знаменитый словарь *COBUILD* был создан, когда корпус уже достиг размера 20 млн слов. Затем из него выросла целая серия разнообразных словарей принципиально нового типа «Ключевые слова в бизнесе» (Mascull 1996), «Ключевые слова в СМИ» (Mascull 1995), «Ключевые слова в науке и технологии» (Mascull 1997) и т.д.

Еще одной инициативой Бирмингемского университета был проект “Bank of English”, начатый в 1991 году с целью составления корпуса из 220 млн слов для отслеживания изменений, происходящих в языке. Предполагалось собрать огромное количество информации без каких-либо предварительно установленных ограничений по объему категорий и постоянно обновлять эту коллекцию по мере поступления нового материала. (Crystal 1995). Сегодня он включает в себя более 500 млн словоупотреблений. В настоящее время исследователям доступен фрагмент Бирмингемской коллекции – Бирмингемский корпус, достигающий 7,3 млн словоупотреблений.

Другим лексикографическим корпусом стал “Longman/Lancaster English Language Corpus”, разработанный в 1980-е годы Деллой

Саммерз (издательство Лонгман) и Джеффри Личем из Ланкастерского университета. Они использовали материалы, опубликованные начиная с 1900 года как на британском, так и на американском вариантах английского языка. К началу 1990-х набралось около 30 млн слов письменного текста и началась работа над устным компонентом корпуса.

“British National Corpus” (Британский Национальный Корпус) стал продуктом сотрудничества между издательствами Longman, Oxford University Press, Chambers Harrap, Компьютерной службой Оксфордского университета, Ланкастерским университетом и Британской библиотекой при поддержке Департамента Торговли и Промышленности. Работа проходила с 1991 по 1994 годы, и результатом стали 100 миллионов слов Британского варианта английского языка – 90 млн в письменных текстах и 10 млн устных – созданных, начиная с 1960 года. Особое внимание уделялось внутреннему балансу корпуса, включая рукописные материалы и такие «несерьезные», обычно не хранимые, вещи, как неформальные записки, напр., молочнику. Данные этого корпуса, представляющие широкий спектр различных видов дискурса, легли в основу последних изданий таких инновационных словарей, как “The Longman Language Activator” и “The Longman Dictionary of Contemporary English”.

В основе популярного словаря “Macmillan English Dictionary” (2002), который впервые включил в словарные статьи информацию о частотности слов и сочетаний, был положен “World English Corpus”.

В конце 80-х гг. под координацией Сидни Гринбаума из Университетского Колледжа Лондона начали создавать “International Corpus of English” с целью представить устные, печатные и рукописные образцы английского языка стран, где он является первым или официальным вторым языком, с тем чтобы каждый национальный компонент состоял из одного миллиона словоупотреблений. К 1991 году 20 стран выразили согласие принять участие в этом проекте, причем некоторые из них запланировали создать национальный региональный корпус. Было дополнительно включено четыре специализированных корпуса: письменные

переводы на английский язык с языков Европейского Союза; устная коммуникация на английском языке между представителями различных национальностей; письменные тексты изучающих английский язык на высоком уровне подготовки; и евро-английский язык (язык, используемый в официальных публикациях Европейского Союза) (Crystal 1997).

Общеизвестны универсальные корпусы большого объема. Такой объем необходим для обеспечения объективной и надежной информации, позволяющей делать обоснованные обобщения. Например, в корпусе, состоящем из миллиона словоупотреблений (что равно по размеру небольшому словарю) представлено только около 50.000 различных слов. Еще одним достоинством таких корпусов является то, что к ним имеют доступ исследователи из разных мест, что позволяет им сопоставлять результаты различных исследований.

Наряду с такими широко известными универсальными корпусами существует множество более мелких, специализированных корпусов. К ним относятся “The Oxford Text Archieve”, представляющий собой коллекцию художественных произведений классической английской литературы; “Corpus of Spoken Professional American English” – собрание пресс-конференций и политических выступлений, проходивших в Белом Доме; “The Michigan Corpus of Academic spoken English” (MICASE), состоящий из образцов исключительно устной и письменной академической речи преподавателей, исследователей и студентов Мичиганского университета (США); “The Bergen Corpus of London Teenage Language”, собранный в 1993 г. и охватывающий 0,5 млн словоупотреблений на базе устной речи лондонских подростков в возрасте от 13 до 17 лет; “Wellington Corpus of Spoken New-Zealand English”; “Penn-Helsinki of Middle English”; “Australian Corpus of English” и многие другие. Список же корпусов, созданных “*ad hoc*”, для какой-либо прагматической цели велик и трудно обозрим.

Долгожданным стало появление в сети Интернет в конце 2004 года Национального корпуса русского языка[41], источником которого являются опубликованные книжные, журнальные и газетные тексты, как правило, в виде выверенных электронных версий, предоставленных издателями или авторами. В этом проекте участвуют специалисты Института русского языка им. В.В. Виноградова РАН, Института языкознания РАН, Института проблем передачи информации РАН, Всероссийского института научной и технической информации РАН и Института лингвистических исследований РАН в Санкт-Петербурге.

На кафедре английского языкознания филологического факультета Московского государственного университета им. М.В. Ломоносова реализуется корпусный проект под руководством проф. Н.Б. Гвишиани[42]. Он представляет собой *The Russian Corpus of Leaner English* – корпуса письменных работ, написанных русскими студентами, изучающими английский язык. Для сопоставления с ним используется корпус *Native Speaker Corpus* аналогичного размера и регистра. Проект позволяет проводить сопоставительные исследования, например, нарушение идиоматичности в текстах русских студентов. В частности, исследование употребления модальных наречий, оканчивающихся на *–ly*, показало, что русские студенты используют исключительно *certainly* и *really* (возможно из-за частотности их русских эквивалентов – *конечно, действительно, безусловно* – которые они автоматически переводят на английский язык). В то время как в английском корпусе используются такие наречия, как *clearly, definitely, apparently* (в русскоязычном корпусе они представлены крайне мало) и *admittedly, presumable, supposedly* (в русскоязычном корпусе отсутствуют полностью).

[41] Подробнее о корпусах русского языка см. Захаров 2014.
[42] См. об этом: Гвишиани 2001.

§ 3.2. Место корпусной лингвистики в системе языковых дисциплин

Корпусная лингвистика представляет собой направление прикладной лингвистики, целью которого является создание и обслуживание (с использованием машинных носителей и соответствующих средств доступа) больших массивов языковых данных, называемых «корпусом».

С помощью корпусов языковых данных можно не только с большой точностью анализировать отдельные факты реализации и использования различных языковых единиц, но и выявлять общие языковые закономерности, уточняя и корректируя ранее выработанные положения различных лингвистических дисциплин. При этом сегодня не вызывает сомнений тот факт, что корпусная лингвистика способствует формированию и открытию новых реалий во многих областях исследования (Гвишиани, Герви 2001).

Лингвистика как наука о естественном человеческом языке вообще и о всех языках мира как индивидуальных его представителях является сама по себе явлением сложным и многогранным. В связи с этим существуют различные попытки деления ее на различные направления и отрасли. Наиболее общим является деление на:

- теоретическую лингвистику (поиск и обобщение научных знаний об устройстве и функционировании языка) и
- прикладную лингвистику (приложение этих знаний в нелингвистических научных дисциплинах и в различных сферах практической деятельности человека, а также теоретическое осмысление такой деятельности).

Наряду с таким широким пониманием термина «прикладная лингвистика» существует и ряд более узких, причем сильно различающихся в разных национальных традициях.

В западной традиции аналоги данного термина (англ. *applied linguistics*, нем. *angewandte Linguistik*) используются прежде всего для обозначения теории и практики преподавания иностранных языков, включая методику, особенности описания грамматики для учебных целей и т.п.

В СССР термин «прикладная лингвистика» получил широкое распространение в 1950-х годах в связи с появлением первых компьютерных систем автоматической обработки текстовой информации (машинного перевода, автоматического реферирования и др.). Именно поэтому в русскоязычной литературе и поныне вместо термина «прикладная лингвистика» в том же значении часто используются термины «компьютерная лингвистика», «вычислительная лингвистика», «автоматическая лингвистика», «инженерная лингвистика», что некоторые ученые (напр., Зубов 1996) считают не вполне удачным, поскольку каждая из перечисленных дисциплин имеет свой предмет и методы работы в рамках прикладной лингвистики как более широкого направления.

Хотя возникновение прикладной лингвистики как автономной научной дисциплины относится к относительно недавнему прошлому (приблизительно к 1920-м годам), прикладные проблемы стояли перед языкознанием практически с самого начала его существования. Эти проблемы, в конечном счете, сводятся к оптимизации функций языка.

С функциональной точки зрения, прикладная лингвистика может быть определена как академическая дисциплина, в которой целенаправленно изучаются и разрабатываются способы оптимизации различных сфер функционирования языковой системы.

В оптимизацию коммуникативной функции вносят вклад такие дисциплины, как:

- теория перевода;
- машинный перевод;
- теория и практика преподавания родного и неродного языка;
- теория и практика информационно-поисковых систем;
- создание информационных и, шире, искусственных языков;
- теория кодирования.

Разнообразие методов прикладной лингвистики связано с разнообразием конкретных областей приложения научных знаний о языке: каждая конкретная прикладная дисциплина обладает своим уникальным набором методов. Тем не менее можно выделить нечто общее, характерное для методов прикладной лингвистики в целом.

Эта общая часть хорошо видна при сравнении методических инструментариев описательной, теоретической и прикладной лингвистики.

Перед описательной лингвистикой стоит задача описания фактов языка. На первом плане при этом находится метод классификации, т.е. выявления той сетки параметров, которая позволяет охватить все существенные свойства языковых структур.

Теоретическая лингвистика формирует само представление о том, какие свойства языка являются существенными, а какие – нет. Создаваемые в теоретической лингвистике концептуальные модели языка не просто описывают наблюдаемые факты, но и претендуют на их объяснение.

Иными словами, классификации языковых фактов и концептуальные модели теоретической лингвистики претендуют на описание того, как действительно устроен язык.

Прикладная лингвистика также использует и метод классификации, и метод моделирования. Однако поскольку задачи прикладной лингвистики сосредоточены в области оптимизации функций языка, а оптимизация определяется конкретной задачей, то в прикладной лингвистике широкое распространение имеет познавательная установка, которая в качестве основной ценности выдвигает не познание того, «как все обстоит на самом деле», а решение конкретной задачи, в частном случае – удовлетворение требований «заказчика», преследующего свои собственные цели. Это, впрочем, не означает, что результаты прикладных исследований не представляют никакой ценности для теории языка: напротив, прикладные модели оказывают значительное влияние на лингвистическую теорию, способствуя обновлению концептуального аппарата современного языкознания.

Суммируя основные частные отличия прикладных моделей от теоретических и описательных, можно сказать, что:

– прикладные модели в целом ориентированы на конкретные подъязыки, а не весь язык в целом;

– они часто (но не всегда) требуют большей степени формализации;
– прикладные модели используют знания о языке выборочно;
– прикладные модели не делают различий между собственно лингвистическими и экстралингвистическими аспектами семантики языковых выражений;
– прикладные модели в существенно большей степени огрубляют моделируемый объект, чем теоретические модели;
– и, наконец, прикладные модели не налагают никаких существенных ограничений на инструмент моделирования.

В наиболее обобщенном виде основные отличия корпусной лингвистики от традиционной лингвистики можно представить следующим образом (приводится по: Рыков В.В. Лекциях по корпусной лингвистике. – http://rykov-cl.narod.ru/c.html):

Таблица 1. Основные отличия корпусной лингвистики от традиционной лингвистики.

Традиционная лингвистика	⇔	**Корпусная лингвистика**
1. Основное внимание – изучение языка	⇔	Основное внимание – изучение речи
2. Цель – описание и объяснение языка	⇔	Цель – описание языка в том виде, как он проявил себя в речи, представленной в виде специально подобранного корпуса текстов
3. В своих исследованиях идет от теории к ее объяснению и подтверждению в фактах речи	⇔	В своих исследованиях опирается на данные корпуса текста
4. Предпочитает квалитативные методы	⇔	Предпочитает квантитативные методы
5. Видит себя частью традиций, базирующихся на рационалистических методах	⇔	Видит себя частью традиций, базирующихся на эмпирических методах

6. Текст рассматривается как некоторая абстракция	⇔	Текст рассматривается как некоторая физическая сущность
7. Рассматривает тексты в локальной перспективе	⇔	Рассматривает тексты в глобальной перспективе
8. Основное внимание – не только форме, но и содержанию	⇔	Основное внимание уделяется форме
9. Изучает языковые универсалии	⇔	Составление грамматики конкретных языков
10.Анализирует некоторую конкретную, искусственно ограниченную, проблемную область	⇔	Фокусирует свое внимание на как можно более широком взгляде на текст, неограниченном ни какими догмами
11.Опирается на интуицию в отборе речевого материала, в отборе эмпирических материалов своих исследований	⇔	В своих выводах опирается на наблюдение речевой деятельности, проявленной в виде текстов
12.Предпочитает логические рассуждения	⇔	Часто пользуется вероятностными методами и статистикой для первичной обработки речевого материала
13.Предпочитаются искусственные примеры, из изолированных от текста словоупотреблений	⇔	Проводится работа с лингвистическими данными (словоупотреблениями) в том виде, в каком они встречались в контексте
14.Предпочитает дедуктивные методы обработки эмпирического словесного материала	⇔	Предпочитает индуктивные методы обработки эмпирического словесного материала, считает их сутью научного метода
15.Верит в открытия, основанные на процедурах, оценках, сравнениях и т.д., т.е., как результат многовековых исследований	⇔	Верит в научные открытия, основанные на обработке эмпирических данных

Неоспоримые достижения в использовании корпусных данных ставят новые вопросы относительно места и роли корпусной лингвистики в иерархии языковедческих дисциплин. Является ли она отдельным лингвистическим направлением или же представляет собой средство (или базу) для проведения лингвистических исследований? Каково ее место и роль в иерархии языковедческих дисциплин? Существуют различные взгляды и мнения на этот вопрос.

Корпусная лингвистика изучает продукт функционирования языка и уделяет основное внимание построению речи в процессе коммуникации. Данное направление возникло как естественное продолжение функционально-коммуникативной парадигмы, получившей практически универсальное признание еще в середине 70-х гг. Успешное функционирование языка в качестве орудия общения обеспечивается взаимодействием языковых элементов, формы которого предопределяются как системой и нормой языка, так и конкретными условиями коммуникации.

По словам Е.С. Кубряковой (1997), «современное состояние теоретической лингвистики характеризуется выдвижением в ней двух главных парадигм научного знания – когнитивной и коммуникативной». Пристальное внимание лингвистов, нацеленное на изучение использования языковых фактов и структур в практике человеческого общения, в процессе построения речи проявилось в обращении к функциональному аспекту языка и послужило почвой для возникновения синтагматических теорий языка, дискурсивного анализа и прагматики.

Однако функциональный подход может быть успешно реализован лишь в тесной связи с подходом структурным. Языковая структура и функционирования должны исследоваться как взаимозависимые и взаимопроникающие аспекты лингвистического анализа.

С точки зрения Н.Б. Гвишиани (Гвишиани, Герви 2001), создание корпуса реального языкового материала, извлеченного из

разнообразных источников и сведенного в компьютеризированную систему, чтобы исследователи могли изучать значения и возникающие языковые закономерности, представляет собой новую стратегию не только в качестве метода, подчеркивающего значимость реальных фактов языка, но и в качестве теории, позволяющей по-новому взглянуть на природу ранее сформулированных положений (Гвишиани, Герви).

Помимо этого, корпусная лингвистика имеет еще две черты, дающие основание претендовать на положение самостоятельной дисциплины:

1. Характер используемого словесного материала.

2. Специфика инструментария.

Таким образом, корпус текстов, с одной стороны, это исходный речевой материал для корпусной лингвистики и для других лингвистических дисциплин; с другой стороны, результат деятельности корпусной лингвистики.

Предметом корпусных исследований выступает продукт, результат деятельности говорящего (или пишущего) в процессе построения речи. Данное многоплановое явление основывается на знании говорящими конкретного языка (*competence*), а также на его/ее «языковом поведении» как члена определенного языкового коллектива и «носителя» определенной лингвистической культуры. В этой связи можно провести параллель с тем, что Ф. де Соссюр определил как *langage* (речевая деятельность) – третий элемент его триады, включающей также *langue* (язык) и *parole* (речь).

Лингвистические данные, как и любая информация, является динамическим объектом, образующимся в момент взаимодействия объективных данных и субъективных методов. По мнению А.В. Всеволодовой (2007), характерной особенностью информации, отличающей ее от других объектов природы и общества, является своеобразный дуализм, проявляющийся в том, что на свойства информации влияют как свойства данных, составляющих ее содержательную часть, так и свойства методов, взаимодействуюших с данными в ходе информационного процесса.

По окончании процесса свойства информации переносятся на свойства новых данных, т.е. свойства методов могут переходить на свойства данных. Это необходимо помнить в связи с тем, что объективные и достоверные на первый взгляд данные могут таковыми не оказаться из-за того, что на отдельных этапах информационного процесса к ним были применены необъективные или неадекватные методы.

Как метод лингвистического анализа корпусная лингвистика связана с контрастивными исследованиями, направленными на установление фактов общего и отличного между языками, диалектами и вариантами языка в ходе их семантического описания, а также на сопоставления устной и письменной речи.

По мнению Н.Б. Гвишиани (2001), «основная цель корпусной лингвистики – исследование языковой действительности, микрокосма языкового использования в процессе коммуникации для наиболее верного и точного определения и демонстрации его различных аспектов».

А.В. Зубов (2004) отмечает, что корпусные данные могут быть использованы для решения следующих лингвистических задач:

- в лексикографии и лексикологии – для составления различных словарей, определения значений многозначных слов, выявления ассоциативных связей слов в тексте, выделения терминов и терминологических сочетаний и т.п.;
- в грамматике – для определения частоты употребления грамматических морфем в текстах различного типа, выявления наиболее употребляемых типов словосочетаний и предложений, определения значений синонимичных морфологических единиц, частоты употребления классов слов и т.д.;
- в лингвистике текста – для дифференциации типов текста, создания конкордансов, выявления связи между предложениями в абзацах и между абзацами и т.д.;
- при автоматическом переводе текстов – для поиска контекстов слов, имеющих несколько переводных эквивалентов, поиска переводных эквивалентов

терминологических и фразеологических словосочетаний в параллельных текстах и т.д.;
- в учебных целях – для выбора цитат, отдельных фрагментов произведений, примеров, используемых в процессе создания учебников и учебных пособий и т.д.

Н.Б. Гвишиани (1997) отмечает еще одно использование корпусной лингвистики – совершенствование педагогической практики по трем основным направлениям:

а) контрастивный анализ, позволяющий заострить внимание обучающихся на трудных элементах иностранного языка, которые обнаруживают значимые структурные различия с родным языком;
б) дискурс-анализ, фиксирующий наиболее частотные, типические и в силу этого «общественно значимые» черты реального речеупотребления;
в) выявление типичных ошибок в грамматике и проявления «неидиоматичности» в развернутом тексте.

Центральным аспектом корпусных исследований представляется языковой выбор – решения, принимаемые говорящим в процессе порождения речи. Проблема языкового выбора ставит нас перед вопросом о том, что в языке можно считать нормальным, общепринятым и что возможным? Здесь корпусная лингвистика пересекается с когнитивными исследованиями, поскольку большинство языковых решений говорящего продиктованы определенной культурой и зависят от нашего знания и понимания мира.

Говоря о языке как таковом, ученые (Гвишиани, Герви 2001) предполагают, что некоторые решения определяются его структурой, другие же являются индивидуальными, единичными. При этом в ряде случаев языковой выбор говорящего отражает определенные тенденции, характеризующие элементы языка прежде всего с количественной стороны, со стороны их употребляемости.

Дискуссии о том, является ли корпусная лингвистика отдельной дисциплиной или методом, привели к формированию двух подходов к корпусным исследованиям[43] – «опирающийся на корпус» (*corpus-based*) и «обусловленный корпусом» (*corpus-driven*) (McEnery et al 2006, Tognini-Bonelli 2001, Bianchi 2012). Сторонники первого используют корпус для иллюстрации, подтверждения или проверки уже существующих лингвистических теорий. Приверженцы второго подхода считают, что сами анализируемые данные ведут к описанию языка или отдельного языкового явления. Теоретически эти подходы восходят к разным источникам – первый к Н. Хомскому (Chomsky 1965), утверждавшему, что информация о языковой компетенции (*competence*) говорящего и слушающего не подлежит прямому наблюдению или извлечению из данных с помощью индуктивных способов. Второй подход основан на контекстуальной теории значения Дж. Р. Ферса (Firth 1957), утверждавшему, что лучше всего понять значение слова можно не изучая его изолированно, а по «компании, с которой оно водится», переходя, тем самым от языковой компетенции к речевой практике (*performance*, по Хомскому или *parole* по Соссюру).

В отношении трех основных свойств корпуса – репрезентативность, объем и аннотированность – два подхода значительно отличаются друг от друга (см. Таблицу 2).

В реальности наряду с такими крайними точками зрения существуют промежуточные, допускающие различные варианты, оговорки и уточнения.

[43] Н.Б. Гвишиани (2016) определяет их следующим образом: корпусные практики, как опирающиеся на заранее предложенную теоретическую концепцию (‘*corpus-based*’), так и исходящие исключительно из статистического анализа данных электронного корпуса (‘*corpus-driven*’).

Таблица 2. Отличия между подходом, основанном на корпусе, и подходе, обусловленным корпусом.

Подход, опирающийся на корпус	**Подход, обусловленный корпусом**
Репрезентативность	
Необходимо соблюдать баланс корпуса с точки зрения темы, стиля, разновидности и пр.	По мере роста корпус достигнет так называемой кумулятивной (естественной) репрезентативности.
Объем корпуса	
«Объем не имеет значения».	«Чем больше, тем лучше».
Аннотированность	
Процесс аннотирования важен сам по себе и основан на существующих теориях.	Против аннотирования. Ученый должен подходить к корпусу, не имея ранее сложившихся теорий, и постулировать языковые категории только исходя из данных корпуса.

1) **Аутентичность данных**. Зачастую то, что люди говорят на самом деле, очень сильно отличается от их представлений о том, что́ они говорят, и еще больше от того, что они, по их мнению, должны говорить. Корпусные данные, созданные и обработанные компьютером, позволяют избежать субъективности и отражают реальную картину объекта.

2) **Репрезентативность** – данные включают в себя не только письменную, но устную речь – от достаточно формальной или, по крайней мере, самоконтролируемой, до случайной, спонтанной, которой в течение долгого времени пренебрегали, считая ее бессистемной, бесформенной, бессвязной. Современные корпуса не ограничиваются материалами художественной литературы (на

которой было построено большинство иллюстративной фразеологии в традиционных словарях и примеров в грамматических справочниках и пособиях), а полноправно включает в себя разнообразные регистры и жанры, как письменной, так и устной речи. Произошел, пользуясь терминологией В.В. Виноградова, поворот от изучения языка художественной литературы к изучению литературного языка.

3) Возможность проведения **количественных исследованиий**. В последнее время стало очевидным, что, например, грамматические системы по своей природе вероятностны; так, например, систему «полярности» в английском языке следует моделировать не просто как противопоставление «положительное или отрицательное», а как «положительное или отрицательное с определенной долей вероятности» (Halliday 2004).

4) **Освобождение от «монополии» интуиции носителей языка**. Использование аутентичного материала корпусов позволяет «не-носителям» английского языка выдвигать свои собственные предположения и гипотезы, находя для них достоверные обоснования, самостоятельно делать выводы и обобщения, не полагаясь всецело на знания и «языковое чутье» лиц, для которых английский является родным языком.

5) **возможность применения** корпуса не только в исследовательских целях, но и **в прикладных, практических целях преподавания английского языка**, что способно сделать обучение более осознанным, интересным, а потому – эффективным.

Дж. Беннет (Bennet 2010) рассуждает о том, чего не может корпусная лингвистика, и приходит к следующему:

– корпусная лингвистика не может дать опровергающее доказательство. Это означает, что корпус не может сказать нам, что возможно и правильно, а что невозможно и неправильно. Многие ошибочно считают, что если какое-то явление не отражено в корпусе, то он не заслуживает доверия. На самом деле, это лишь означает что данное явление достаточно редко встречается в регистре, представленном в корпусе;

– корпусная лингвистика не может дать ответ на вопрос *почему*? Она лишь может дать ответ на вопрос *что*? Для объяснения причин потребуется интуиция пользователей языка;

– корпусная лингвистика не может охватить весь язык сразу. Независимо от того, насколько тщательно был спланирован корпус, от его объема (напрмер, *Cambridge International Corpus*, состоящий из миллиарда слов) он не может отразить все явления языка.

В качестве проблемы корпусных исследований можно отметить тот факт, что лингвисты, специализирующиеся в них, часто являются простыми «собирателями данных». Но эти данные ничего не прибавят в нашем понимании языка, если они не будут обрабатываться и осмысливаться, исходя из запаса имеющихся теоретических знаний.

В связи с этим представляется обоснованным подход, предложенный Т.Б. Назаровой (2012) и основанный на необходимости сочетания корпусного подхода с когнитивным, который «настаивает на признании роли мыслительных процессов, подчеркивает необходимость осознанного и осмысленного управления ими, демонстрирует неразрывную связь понятийного и языкового развития, внедряет методологический принцип единства языка и мышления, языка и культуры».

4. Основы корпусного анализа

§ 4.1. Типология корпусов

Термин «корпус» в наши дни получил широкое распространение, однако понимается зачастую по-разному – от любого массива текстов до массива машиночитаемых (электронных текстов) и, наконец, собрания машиночитаемых текстов, имеющего конечный объем и отобранных так, чтобы максимально полно отражать определенную разновидность языка. Поэтому, несмотря на огромное количество определений[44] данного термина (см., например, Baroni 2009, Biber et al 1998, Gries 2009, Hunston 2006, McEnery 2003, Meyer 2002 и многие другие), наблюдается, как отмечает Тони МакЭнери (McEnery et al. 2006), растущий консенсус в отношении того, что корпус – это собрание: (1) машиночитаемых (2) аутентичных текстов (включая транскрипцию устных текстов), отобранные так, чтобы быть (3) представительными в отношении того или иного языка или разновидности языка.

Существуют различные подходы к классификации корпусов текстов в зависимости от типа текстов, способов их организации, языка и т.д. (Sinclair 1996, McEnery, Wilson 2001, Leech 1991). Рассмотрим некоторые из них.

---

[44] В словаре корпусной лингвистики *A Glossary of Corpus Linguistics* (Baker et al. 2006) приводится определение корпуса, предложенное Сьюзан Ханстон (Hunston 2002:2-3): от латинского *corpus* (мн. ч. *corpora*) – «тело» – собрание текстов («тело» языка) хранящееся в электронной базе данных. Корпусы обычно представляют собой большие массивы машиночитаемых текстов, содержащих тысячи или миллионы слов. Корпус отличается от архива тем, что часто (хотя и не всегда) тексты отбираются так, чтобы их можно было назвать представительными для определенной разновидности языка или жанра, и, следовательно, могут использоваться в качестве образца для сравнения. Корпусы часто сопровождаются дополнительной информацией, например, метки, указывающие на часть речи (POS) или особенности интонации для устной речи. Отдельные тексты внутри корпуса могут в заголовке иметь закодированную информацию о жанре, авторе, дате публикации и т.д. Корпусы могут быть быть специализированными, справочными, многоязычными, параллельными, учебными, диахроническими и мониторными. Хотя сам корпус не содержит новой информации о языке, с помощью программного обеспечения для обработки данных можно получить новый взгляд на уже известное.

I. По степени охвата языкового материала:

1) **Общие / национальные / стандартные / справочные** корпусы (*general / national / standard / sample / / core /reference*). К ним относятся корпусы первого поколения 1960х – 1970х годов (*Brown Corpus of written American English, Lancaster Oslo-Bergen (LOB) corpus of written British English* и др.), заложившие стандарт (500 текстов по 2000 слов в каждом из различных письменных жанров), по которому стали строиться новые корпусы, что сделало возможным использовать их для сравнения языкового материала. Особенностью второго поколения (1990-е годы), которое стало возможным благодаря распространению сети Интернет, стал огромный объем корпусов. Первым мега-корпусом (1991) стал *Bank of English*, содержащий не только письменные, но и устные тексты, за ним последовал амбициозный проект *British National Corpus,* который уже в 1994 году насчитывал 100 миллионов слов, а затем и американский аналог *Corpus of Contemporary American English***.** Эти корпусы дают исследователю картину «языка в целом», поэтому, чем значительней их объем и и чем больше разновидностей языка они представляет, тем более полной является эта картина.

2) **Специализированные корпусы** (*specialised*) не ставят перед собой целью представлять язык в целом, они ориентированы на отдельные сегменты (области, жанры). Поэтому они почти всегдя меньше по объему, чем общие корпусы, именно в силу более узкого фокуса. Репрезентативность такого корпуса достигается за счет большей однородности текстов данной области, а не за счет его объема. Причем, как отмечает Сьюзан Ханстон (Hunston 2002), не существует границ специализированности такого корпуса, устанавливаются лишь параметры, ограничивающие типы входящих в него текстов. Корпус может быть ограничен временными рамками (одним веком), социальной средой (например, разговоры в книжном магазине) или темой (газетные статьи, посвященные Европейскому Союзу). Актуальными в наши дни специализированными областями являются **академическое общение** (*Michigan Corpus of Academic Spoken English* (США),

Limerick-Belfast Corpus of Academic Spoken English (Ирландия) и *City University Corpus of Academic Spoken English* (Гонконг) и др.), **деловое общение** (*Wolverhampton Business English Corpus* и *Business Letters Corpus*), а также **профессиональное общение** (*Professional English Research Consortium, Air Traffic Control (ATC) Corpus, HKUST Computer Science Corpus, Corpus of Professional Spoken American English* и многие другие).

3) **Индвидидуальные / самодельные / импровизированные корпусы** (*homemade / adhoc/ DIY*) – непрофессиональные корпусы, созданные для узких исследовательских или педагогических целей, для конкретного небольшого проекта, обычно небольшого размера, в основном, состоящий из письменных текстов.

II. По способу применения корпуса (Зубов 2004):

1) **Исследовательские корпусы** – создаются с целью изучения различных аспектов функционирования языка. Они строятся до проведения какого-либо исследования. Как правило, такие корпусы текстов содержат несколько десятков миллионов словоупотреблений.

2) **Иллюстративные корпусы –** служат для выделения из них лингвистических примеров, подтверждающих те или иные языковые (речевые, текстовые) факты, обнаруженные ранее иными лингвистическими приемами. Они создаются после проведения: их цель не столько выявить новые факты, сколько подтвердить и обосновать уже полученные результаты. Такие корпусы не являются слепком, правильным (с точки зрения статистики) отображением проблемной области.

3) **Параллельные корпусы** – состоят из множества текстов-оригиналов, написанных на каком-либо исходном языке, и текстов-переводов этих исходных текстов на один или несколько других языков. Они дают ценный материал для проведения сравнительно-сопоставительных исследований и для обучения переводу человека и компьютера.

Иногда такого рода корпусы называются не «параллельными», а «переводными» (*translation corpus*), а под собственно «параллельным корпусом» (*parallel corpus*) подразумевают тематически сопоставимые оригинальные тексты на двух и более языках. Если первый тип сопоставления помогает лучше понять природу языка перевода, то второй – служит целям контрастивного анализа (Гвишиани 1997). Кроме того, переводные корпусы бывают «**выравненными**» (*aligned*), когда два текста могут анализироваться по соответствующим сегментам. Размер таких сегментов может варьироваться, но обычно равен предложению.

III. По хронологическому признаку:

1) **Статические / синхронические корпусы** (*static / synchronic*) – содержат тексты какого-то небольшого временно́го промежутка. Первоначально корпусы текстов создавались как статические образования. При интерпретации тех или иных языковых явлений следует соблюдать осторожность, потому что корпус может пополняться и изменяться с течением времени. Статический корпус позволяет избежать такой опасности. Среди самых крупных и известных статических корпусов можно назвать *International Corpus of English, Longman/Lancaster Corpus, Longman Written American Corpus*. Типичными представителями этого вида корпусов являются авторские корпусы – коллекции текстов писателей, например, корпус пьес Шекспира (*Shakespare Corpus*).

2) **Динамические / мониторные корпусы** (*dynamic / monitor*) – объем таких корпусов постоянно пополняется и потому неограничен. Говоря словами Джона Синклера (Sinclaire 1991:25), «они постоянно развиваются, как и сам язык». Они преимущественно используются в лексикографии, а также для произведения различных диахронических исследований, выявления исторических изменений в функционировании языковых явлений – например, изменение значений слов, частоты использования тех или иных синтаксических конструкций и т.д.

Для этого была разработана специальная технология построения и эксплуатации динамического корпуса текстов (Баранов 2001). В имеющейся литературе такие корпусы получили название мониторных. Особенность сборки мониторных корпусов заключается в том, что они не предполагают раз и навсегда заданного набора текстов. В течение заранее зафиксированного промежутка времени происходит обновление и/или дополнение множества текстов корпуса. Самым известным мониторным корпусом является *Bank of English*, начатый в 1991 году на базе проекта *COBUILD (Collins Birmingham University International Language Database*). К этому типу корпусов также относится *Global English Monitor Corpus* (начатый в 2001 году), а также многочисленные корпусы «брауновского семейства», построенные по образцу Брауновского корпуса (*Brown University Standard Corpus of Presentday American English* (Kučera, Francis 1967) – *Frown, FLOB, ACE (Australian Corpus of English* известный также как *Macquarie corpus), WWC (Wellington Corpus of Written New Zealand English), Kolhapur (the Kolhapur Corpus of Indian English)* и многие другие. Недостатком таких корпусов ученые McEnery, Wilson 1996) считают невозможность проведения сопоставительных исследований из-за постоянно меняющегося состава корпуса.

3) **Диахронические / исторические корпусы** (*diachronic / historical*) – отдельный вид корпусов, состоящих из текстов одного языка, принадлежащих к разным периодам времени (гораздо более продолжительных, чем в динамических мониторных корпусах) и использующихся для изучения эволюции языка. К числу диахронических текстов относятся: *Helsinki Corpus of English Texts, A Representative Corpus of Historical English Registers ARCHER, Lampeter Corpus of Early Modern English Tracts, Corpus of Early English Correspondence, Dictionary of Old English Corpus in Electronic Form*, базы данных *EEBO (Early English Books Online)* и ECCO *(Eighteenth Century Collections Online).*

IV. По представленности языкового материала:

1) **Полнотекстовый корпус** – содержащий тексты целиком (например, *British National Corpus*).

2) **Фрагментированный корпус** – содержащий определенным образом ограниченные фрагменты текстов, что обеспечивает сбалансированность собранного материала. Примером может служить Брауновский корпус (*Brown University Standard Corpus of Present-Day American English, Brown Corpus*), состоящий из 500 фрагментов объемом около 2000 слов из текстов 15 разных жанров. При этом доля фрагментов одного жанра соответствует доле всех опубликованных текстов этого жанра. По замыслу составителей, именно это обеспечило сбалансированность собранного материала.

V. По языку:

1) одноязычный,

2) двуязычный,

3) многоязычный.

V. По степени организации и структурированности (Atkins et al., 1992):

1) **Электронный архив** (*archive*) – тексты на электронном носителе, но их форма, представленная на машинном носителе, не стандартизирована и не унифицирована (напр., *Oxford Text Archive*).

2) **Электронная библиотека** (*electronic text library*) – тексты представлены однородным и стандартизированным образом, но без четких ограничений по отбору материала.

3) **Корпус текстов** (*corpus*) – форма стандартизирована и унифицирована, тексты предназначены для отражения части лингвистической реальности (напр., *Oxford Pilot Corpus*).

4) **Субкорпус / подкорпус** (*subcorpus*) – некоторая автономная часть корпуса.

VI. По типу носителя источника:

1) **Письменный корпус** (*written*) – состоит только из текстов, созданных или опубликованных в письменном виде. Сюда относятся традиционные книги, учебники, газеты, журналы или неопубликованные письма и дневники, а также электронные тексты, такие как электронная почта, блоги, Интернет-сайты и т.п. В отношении того, что следует относить к письменным текстам, возникает спорный вопрос – относятся ли, например, заранее подготвленные речи, сценарии теле– и радиопередач, фильмов к текстам, которые можно охарактеризовать как «написанные для последующего говорения». При отсутствии электронных версий тексты обычно набираются на компьютере или сканируются с последующим распознаванием. К письменным относятся корпусы «брауновского семейства» (*Brown University Standard Corpus of Present-Day American English, Brown Corpus)*, исторические корпусы, а также многочисленные специализированные корпусы, созданные из письменных источников.

2) **Устный корпус** (*spoken / speech*) – состоит только из транскрипции устных текстов, таких как бытовой разговор, совещания, лекции, радиопередачи и т.п., например, *Spoken English Corpus, London-Lund Corpus of Spoken English (LLC), Cambridge and Nottingham Corpus of Discourse in English (CANCODE), Switchboard Corpus, the Freiburg Corpus of English Dialects (FRED), Michigan Corpus of Academic Spoken English (MICASE)* и многие другие.

3) **Смешанный корпус** (*mixed*) – состоит из письменных и устных текстов (*Bank of English, British National Corpus* и многие другие).

4) **Мультимедийный корпус** (*multimedia / multimodal*) – возникающие в последнее время корпусы имеют письменную расшифровку устных текстов, которая синхронизирована и выровнена с исходной аудио– или видеозаписью. Исследователи могут выйти за рамки расшифровки и охватить все дискурсивные элементы: интонацию, жесты, взгляд, расстояние и т.п. Некоторые такие корпусы могут запускаться и просматриваться через веб-

браузер – на сайте *TalkBank* можно воспроизводить аудио- и видеофайлы корпуса *Santa Barbara Corpus of Spoken American English (SBCSAE)* с прокруткой расшифрованного текста. В качестве полезного ресурса при обучении английскому языку можно привести мультимедийный корпус *ELISA (English Language Interview Corpus as a Second-Language Application)*, который представляет собой небольшое собрание интервью с людьми разных профессий, сопровождаемые мультимедийными средствами, а также упражнениями и другими учебными материалами. Как отмечает Сабине Браун (Braun 2005), такая интеграция теста и видео способствует контекстуализации материала для его лучшего усвоения.

VII. По индексации:

1) **Простой / неаннотированный / неразмеченный корпус** (*raw / unannotated*) – собрание текстов в электронном виде без предварительной обработки в виде включения дополнительной информации о составляющих его языковых данных. Иногда к этому типу относят корпусы, содержащие **метаданные** о входящих в него текстах (название, источник, время создания, автор и т.п.)

2) **Аннотированный / размеченный / таггированный / индексированный корпус** (*annotated / tagged / marked-up*) – корпус, прошедший обработку (вручную или автоматически), заключающуюся в присвоении языковым данным определенных буквенных или цифровых кодов (индексов), которые обозначают их грамматические, лексические, семантические или структурные признаки. Некоторые авторы (Lee 2001) разграничивают понятия «аннотирование» и «разметка». Под **разметкой** (таггированием) имеется в виду добавление цепочек символов для кодирования структурных или поверхностных атрибутов текста (напр., заголовки, разделы, разрывы страниц, предложения, полужирный шрифт / курсив, обозначение говорящего, очередность говорящих, паузы), а также неинтерпретируемых аспектов ситуационного контекста дискурса (напр., библиографические или

демографические сведения об авторе или говорящем, место действия, жанр и т.п., а также жесты, смех, особенности голоса и такие действия как «пишет на доске» и пр.). В языках разметки HTML/SGML/XML разметка расположена между косыми чертами. В свою очередь, **аннотация** – это подмножество разметки; тэги (добавленные цепочки знаков), используемые для кодирования аргументированной или интерпретируемой информации, полученной в результате анализа человеком или машиной обычно для исследовательских целей. Наиболее распространенными видами аннотаций являются частеречная, морфологическая, семантическая, прагматическая, дискурсивная и т.д. Первым корпусом, получившим автоматическое аннотирование был Брауновский корпус (*Brown University Standard Corpus of Present-Day American English, Brown Corpus).* Более подробно аннотирование будет рассмотрено в следующем параграфе.

В последние годы было создано огромное количество корпусов – от обширных национальных корпусов до более узких специализированных[45]. Многие из них доступны для всех желающих (бесплатно или за определенную плату), что позволяет осуществлять поиск тех или иных слов или языковых явлений, а также использовать их для сравнения, например, с данными своего корпуса. В следующей таблице перечислены лишь несколько наиболее известных корпусов каждого типа.

[45] Актуальные перечни корпусов и информацию о них можно найти в каталогах CLARIN (www.clarin.eu/) и ELRA (http://www.elra.info/), а также на сайте Ассоциации Association for Computational Linguistics (https://aclweb.org/aclwiki/List_of_resources_by_language).

Таблица 3. Наиболее известные типы корпусов.

Название корпуса	Информация о корпусе
Исследовательские национальные корпусы (динамические)	
British National Corpus	**Разработчик**: Oxford University Press, Longman, British Library **Время создания**: 1991-1995 **Объем**: 100 млн слов **Содержание**: сбалансированный корпус современного британского варианта английского языка: 4124 текста разных жанров (90% письменных, 10% устных). **Разметка**: есть **Доступность**: бесплатный поиск по корпусу http://www.natcorp.ox.ac.uk
BROWN Corpus	**Разработчик**: Н.Френсис и Г.Кучера, университет Брауна, г. Провиденс, штат Род Айленд, США **Время создания**: 1960-е гг. **Объем**: 1 млн слов **Содержание**: 500 письменных текстов 15 категорий объемом 200 слов **Разметка**: в некоторых версиях (частеречная) **Доступность**: компакт-диск международной организации компьютерной лингвистики ICAME (International Computer Archive of Modern and Medieval English) http://clu.uni.no/icame/
American National Corpus	**Разработчик**: консорциум, состоящий из Oxford University Press, Cambridge University Press, издательства Langenscheidt и корпорации Microsoft **Время создания**: начало – 1998 г. **Объем**: в 2003 г. – 11 млн слов. Планируется – 100 млн. **Содержание**: письменные и устные тексты американского варианта английского языка, включая современные жанры, такие как интернет-блоги и т.п. **Разметка**: есть, разнообразная **Доступность**: часть корпуса OANC (Open American National Corpus) доступна http://www.anc.org/

	Исторический (диахронический корпус)
Bank of English (Collins Cobuild)	**Разработчик**: Бирмингемский университет (Дж.Синклер и его група) и издательство Harper-Collins **Время создания**: начало – 1980-е гг. **Объем**: 650 млн слов, пополняется. Является частью Collins Corpus, насчитывающего более 4,5 млрд слов. **Содержание**: 75% письменных и 25% устных текстов британского (70%), американского (20%) и других вариантов английского языка. **Разметка**: есть (частеречная) **Доступность**: при регистрации http://www.cqpweb.bham.ac.uk/usr/?thisQ=create&uT=y
Corpus of Contemporary American English (COCA)	**Разработчик**: Марк Дэвис, проф. Кафедры корпусной лингвистики университета Бригама Янга, штат Юта, США **Время создания**: начало – 1990. **Объем**: 560 млн слов (каждый год пополняется на 20 млн слов). **Содержание**: равномерный баланс 5 типов текстов: устная речь, художественная литература, популярные журналы и газеты, научные журналы. Тексты взяты из сети Интернет и других электронных ресурсов, поэтому возможен только поиск, но не загрузка всего корпуса. **Разметка**: есть (частеречная) **Доступность**: бесплатно для ограниченного числа запросов в день, неограниченный доступ и доступ к текстам – за плату https://corpus.byu.edu/coca/
Национальный корпус русского языка	**Разработчик**: Институт русского языка имени В.В. Виноградова РАН, Институт языкознания РАН, Институт проблем передачи информации имени А.А. Харкевича РАН, Институт лингвистических исследований РАН в Санкт-Петербурге (совместно с Санкт-Петербургским государственным университетом), Воронежский государственный университет. **Время создания**: начало – 2004, пополняется. **Объем**: 283 млн слов (основной корпус), 600 млн (общий корпус). **Содержание**: письменные тексты (художественные, мемуары, публицистика, научная, религиозная литература, повседневная печатная продукция), записи устных текстов (публичной речи и частных бесед). **Разметка**: есть **Доступность**: В настоящее время свободным и бесплатным является только поиск по корпусу. Доступ ко всему корпусу невозможен в связи с законом об авторских правах. Для получения доступа к 1/6 размеченной части подкорпуса

	необходимо зарегистрироваться и принять лицензионное соглашение http://www.ruscorpora.ru/
	Корпусы устной речи
London-Lund Corpus of Spoken English (LLC)	**Разработчик**: Р. Кверк и С.Гринбаум из University College London (Великобритания) и Й.Свартвик из Lund University (Швеция) **Время создания**: 1960-е, 1975-81, 1985-88 Основан на двух проектах Survey of English Usage (SEU, 1959, University College London) и Survey of Spoken English (SSE, 1975, Lund University) **Объем**: 500 млн слов. **Содержание**: устная речь (бытовые разговоры – непосредственные и по телефону) и публичные обсуждения (спонтанные и подготовленные). **Разметка**: есть (просодическая и дискурсивная) **Доступность**: компакт-диск международной организации компьютерной лингвистики ICAME (International Computer Archive of Modern and Medieval English) http://clu.uni.no/icame/
	Параллельный корпус
The United Nations Parallel Corpus v1.0	**Разработчик**: Организация Объединенных Наций **Время создания**: 1990-2014 **Объем**: 98 млн слов (19 млн на каждый язык) **Содержание**: официальные отчеты и и другие документы заседающих органов Организации Объединенных Наций на всех шести официальных языках ООН: испанском, русском, французском, английском, китайском (мандарин) и арабском **Разметка**: каждому документу, преобразованному в файл в формате XML, присваиваются метаданные (номер документа, номер перевода, место публикации и т.п. **Доступность**: льготная лицензия http://conferences.unite.un.org/UNCorpus
	Сопоставительный корпус
International Corpus of English	**Разработчик**: С. Гринбаум, сейчас координатор – Дж. Нельсон, кафедра английского языка и литературы Университетского коледжа Лондона (Великобритания). У каждого подкорпуса – свой разработчик **Время создания**: 1990 **Объем**: 98 млн слов (19 млн на каждый язык) **Содержание**: каждый подкорпус состоит из 500 текстов (300 письменных и 200 устных) по 2000 слов каждый. В конечном итоге общий объем должен составить 1 млрд слов и включать такие варианты английского языка как британскай,

	восточноафриканский, индийский, новозеландский, филиппинский, сингапурский, американский, австралийский, канадский, гонконский, ирландский, ямайский, малазийский, южноафриканский и шри-ланкийский **Разметка**: да **Доступность**: некоторые подкорпусы – на компакт-диске ICAME, другие – свободно http://ice-corpora.net/ice/avail.htm
International Corpus of English	**Разработчик**: М. Риссанен, О. Ихалайнен, М. Кютоо, кафедра английского языка университета Хельсинки. **Время создания**: 1984-1991 **Объем**: 1.6 млн слов **Содержание**: структурированный многожанровый корпус, состоящий из 450 древнеанглийский, среднеанглийских и ранненовоанглийский текстов (период с 730 по 1710 гг). **Разметка**: да (информация о каждом тексте) **Доступность**: на компакт-диске ICAME http://clu.uni.no/icame/
	Учебный корпус
International Corpus of Learner's English	**Разработчик**: С. Грэйнджер, университет Лувен-ла-Нев, Бельгия **Время создания**: 1990-2000 **Объем**: 2,5 млн слов **Содержание**: состоит из эссе, написанных студентами, изучающими английский яык на продвинутом уровне. Объем каждого эссе – 500-1000 слов. Студенты представлены следующими странами: Болгария, Чехия, Голландия, Финляндия, Франция, Германия, Италия, Польша, Россия, Испания и Швеция. **Разметка**: нет **Доступность**: компакт-диск http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm
	Специализированные корпусы
Michigan Corpus of Academic Spoken English (MICASE)	**Разработчик**: Дж.М.Суэйлз, Р.Симпсон, С.Бриггс, Дж. Овенс, Институт английского языка Мичиганского университета (США) **Время создания**: 1997 (продолжается) **Объем**: 1,8 млн слов **Содержание**: транскрипция 200 часов аудиозаписей общения в университетской среде (лекции, обсуждения на занятиях, лабораторные занятия, семинары, консультации и т.п.) **Разметка**: да (дискурсивная) **Доступность**: да

	https://quod.lib.umich.edu/cgi/c/corpus/corpus?page=home;c=micase;cc=micase
Wolverhampton Business English Corpus	**Разработчик**: группа компьютерной лингвистики Университета Вулвергемптона (Великобритания) по программе ELRA Европейской комиссии **Время создания**: 1999-2000 (6 месяцев) **Объем**: 10 млн слов **Содержание**: тексты 23 Интернет-сайтов, включая описания продукции, пресс-релизы компаний, ежегодные финансовые отчеты, деловая пресса, научные статьи и т.п. **Разметка**: да **Доступность**: платный http://catalog.elra.info/en-us/repository/browse/wolverhampton-business-english-corpus/27ca822ca9db11e7a093ac9e1701ca023f771747e03c49b9b7f21c37fe62cc71/
Professional English Research Consortium (PERC)	**Разработчик**: Консорциум Professional English Research Consortium **Время создания**: 1995-2002 **Объем**: 17 млн слов **Содержание**: тексты научных журналов из базы данных JCR (по 50000 слов на каждый журнал) в области естествознания, машиностроения, телекоммуникаций и т.п. **Разметка**: да **Доступность**: бесплатный (при регистрации) до августа 2018 г. http://scn.jkn21.com/~perc04/

§ 4.2. Контекстная информация и лингвистическая аннотация

Если текст представлен в «сыром», необработанном виде, т.е. в виде собрания текстов на электронном носителе, то с помощью специальных компьютерных программ можно получить информацию о количестве слов в корпусе, о наиболее частотных словах, сочетаемости слов, частотных сочетаниях и т.п. Однако многие лингвистические задачи с их помощью не могут быть решены. Так, во многих языках нельзя установить принадлежность слова к тому или иному грамматическому классу, что, в свою очередь, не позволяет определить частоту употребления грамматических классов слов, структуры предложений на уровне классов слов (частей речи), а значит, и употребительность таких структур. Без специальной подготовки текста не разрешимы омография морфем и слов, полисемия и целый ряд других важных задач.

Это вызвало необходимость создания таггированных (размеченных) корпусов текстов (от англ. *tag* – «индекс, помета»). Все слова такого корпуса получают некоторые буквенные или цифровые индексы, которые обозначают их грамматические, лексические, семантические или структурные признаки.

Как уже отмечалось в предыдущем параграфе, существуют разные подходы к разметке текстов и их единиц в корпусе. Даже сами термины разметка и аннотирование по-разному спользуются и интерпретируются разными авторами. Одни считают их равнозначными, другие – различают их. Обычно принято считать, что разметка бывает двух типов – **разметка документа** (*document markup*) и **аннотирование** (*annotation*).

Разметка документа похожа на разметку с помощью кодов HTML, которая используется для создания документов в сети Интернет (веб-страниц). С помощью этих кодов создается веб-страница, содержащая все нужные элементы: на ней можно

разместить простой текст, выделить его жирным шрифтом или курсивом, вставить ссылку, таблицу, нумерованный или ненумерованный список, картинки, разбить текст на абзацы и разделы, дать разделам заголовки.

Помимо таких «технических» данных разметка содержит метаданные – данные о данных. Будучи эмпирической наукой, корпусная лингвистика не может обойтись без дополнительнй информации относительно того или иного произведения речи. Метаданные могут восстановить его контекст и помочь поместить тот или иной образец речи в его «естественную среду обитания» (Burnard 2004). В число метаданных входят следующие группы:

– редакционные метаданные, предоставляющие информацию об отношениях между компонентами корпуса и их оригинальным источником;
– аналитические метаданные, предоставляющие информацию о том, каким образом компоненты корпуса интерпретировались и анализировались;
– описательные метаданные, предоставляющие классифицирующую информацию внутренних и внешних свойств компонентов корпуса;
– административные метаданные, предоставляющие документальную информацию о самом корпусе – его название, доступность, редакционная версия и т.п.

Изначально, в первых корпусах, таких как Брауновский или *LOB*, эта информация содержалась в объемных бумажных руководствах по использованию. Теперь все метаданные интегрированы в сам корпус с использованием унифицированных принципов кодирования и языков разметки, что облегчает автоматическую оценку точности и единообразия документации, упрощает разработку удобного для пользователя программного обеспечения для доступа к данным и помогает обеспечить соответствие корпуса и метаданных, которые могут распространяться как единое целое.

Важным этапом на этом пути было создание консорциума по кодированию текстов – *Text Encoding Inititiative (TEI)*[46], которая в 1994 году впервые опубликовала объемное руководство по кодированию машиночитаемых данных (*Guidelines for the Encoding of Machine Readable Data*). Эти рекомендации, которые впоследствии обновлялись и дополнялись, получили широкое признание и ипользуются для стандартизации языковых ресурсов. Ключевым компонентом рекомендаций стало определение особого компонента метаданных, известного как *TEI Header* – закодированный особым образом заголовок каждого документа.

4.2.1. Контекстная информация (разметка)

Описательные данные содержат в себе социальный контекст (время, место и участники коммуникации), в котором возник каждый из речевых образцов, составляющих корпус, что многие ученые (напр., Burnard 2004) считают столь же важным (если не более важным), чем его собственно языковые свойства.

Если для письменного корпуса информацию (автор, источник, время создания, издательство и т.п.) о входящих в него текстах можно восстановить, поскольку письменные тексты существуют сами по себе и имеют свою историю, то для устных текстов информацию о месте и времени записи, демографических данных говорящего и слушающего, ситуации и обстановке общения, которая представляет огромную важность для исследователя, найти практически невозможно.

Маркировка типа текста или жанра, используемая в том или ином корпусе, может быть почерпнута из открытого множества, допускающего изменения, однако удобно брать ее из заранее определенного набора значений (таксономии). Иногда используются оба подхода – например, в Британском национальном корпусе (*BNC*) каждый текст связан с открытым множеством дескриптивных ключевых слов, относящихся к области, которой принадлежит текст,

[46] Подробнее о консорциуме TEI можно узнать на сайте http://www.tei-c.org/

а также с набором заранее определенных кодов области. Так, текст *B1G* корпуса *BNC baby corpus*[47], который представляет собой отрывок из книги по геоинформатике, содержит в заголовке, среди прочего, следующую информацию:

```
<catRef target="alltim3 acad  wriase0  wridom3   wrista2 "/>
    <classCode scheme="DLee">W ac soc science</classCode>
    <keywords scheme="COPAC">
      <term>Geography – Methodology – Addresses, essays, lectures</term>
      <term> Geographical information systems.</term>
      <term> Geography – Computer programs</term>
    </keywords>
```

Первая строка указывает, к какому классу относится текст согласно классификации для всего корпуса и состоит из серии кодов (`alltim3, acad` и т.п.), каждый из которых далее определяется в заголовке корпуса. Вторая строка указывает, к какому классу относится текст, согласно классификации Дэвида Ли (2001), разработанной им для Британского национального корпуса в целом, и также использует заранее определенные коды, такие как `W` (письменный), `ac` (академическая проза) и т.д. В оставшейся же части показано, к какому классу принадлежит исходный текст согласно британскому библиотечному каталогу COPAC с использованием каталожных дескрипторов.

Существуют разные подходы к разметке текста и его метаданных. Наиболее распространенными являются следующие атрибуты текста:

а) название текста (полное и краткое),

б) фамилия, имя автора,

в) псевдоним автора (если есть),

г) дата рождения автора (и смерти, если автор умер),

д) время создания текста (или его первой публикации),

---

[47] Один из двух подкорпусов (наряду с *BNC Sampler*) Британского национального корпуса (*BNC*), состоящий из четырех наборов образцов по миллиону слов каждый. Слова в каждом наборе соответствуют конкретной жанровой категории. Один набор образцов содержит транскрипции разговоров, а остальные три набора содержат образцы письменных текстов из научной литературы, художественной литературы и газет.

е) место первой публикации),
ж) тип текста (художественный, научный и т.п.),
з) графика текста (латиница, кириллица и т.п.),
и) наличие в тексте таблиц, графиков, рисунков,
к) указание на рецензии и критические материалы к тексту,
л) адрес места хранения текста,
м) другая информация, важная для проведения конкретных исследований.

Классификации подвергаются и текстовые компоненты. Например, при расшифровке и транскрибировании устного текста полезно указывать, к какому конкретно лицу относится высказывание, чтобы можно было отличить речь мужчин и женщин или представителей разных социальных или экономических групп. Ключевым моментом здесь является наличие средств, позволяющих записывать информацию об отдельных людях один раз для всех в заголовке текста. Для каждого говорящего определяется набор элементов, содержащих такие переменные как возраст, социальное положение, пол и т.д., которые затем группируются внутри элемента <person>, например:

```
<person id="S1">
  <occupation>student</occupation>
  <sex>male</sex>
  <ageGroup>15-20</ageGroup>
</person>
<person id="T3">
  <occupation>instructor</occupation>
  <sex>female</sex>
  <ageGroup>30-35</ageGroup>
</person>
```

Затем внутри текста для каждого высказывания можно определить говорящего с помощью идентифицирующего кода, представленного как величина для атрибута **id**, как показано в вышеприведенном примере:

```
<u who="T3">Good morning class</u>
<u who="S1">I didn't do it</u>
```

Атрибута `who`, которым снабжен каждый элемент `<u>`, достаточно, чтобы определить, какой говорящий имеется в виду. Чтобы отобрать высказывания по говорящим в соответствии с определенными критериями (напр., все высказывания мужчин, все высказывания тренеров определенной возрастной группы) используется эквивалент соединения высказывания и говорящего с использованием значения идентификатора. Этот метод упрощает кодирование текста, поскольку нет необходимости давать информацию о, скажем, поле или возрасте для каждого высказывания, что увеличивает объем информации. Если для каждого говорящего появляется новая категория информации, ее нужно лишь добавить к элементу `<speaker>`, чтобы потом использовать в запросах по всем существующим группам корпуса. Этот же метод используется для отбора высказываний, относящихся к конкретной социальной ситуации или обстановке.

4.2.2 Лингвистическая аннотация

Если разметка дает относительно объективную и верифицируемую информацию о компонентах корпуса и структуре каждого текста, то аннотирование имеет дело с интерпретируемой лингвистической информацией. По определению Джеффри Лича (Leech 1997), аннотирование – это процесс добавления такой интерпретируемой лингвистической информации к электронному корпусу устных и/ или письменных языковых данных.

Одним из наиболее распространенных видов аннотаций является добавление к каждому слову в тексте тэгов[48] (дескрипторов) с указанием на часть речи (так называемая частеречная аннотация, *POS tagging*). Такой вид аннотаций полезен для различения слов, имеющих одинаковое написание, но принадлежащих к разным частям речи. Например, слово *present* в английском языке может быть существительным (*презент, подарок*), глаголом (*презентовать, дарить*) и прилагательным (*присутствующий*). С помощью частеречного таггирования это слово может быть аннотировано следующим образом:

`present_NN1` (нарицательное существительное в ед. числе)
`present_VVB` (базовая форма смыслового глагола)
`present_JJ` (качественное прилагательное)

Говоря об «интерпретируемой» информации, Дж. Лич (1997) отмечает, что аннотация всегда в какой-то степени является

[48] Тег (от англ. *tag* – ярлык, бирка) – именованная метка, более правильное название – дескриптор. Это элемент языка программирования, необходимый для разметки гипертекста, например: <big> - <big>. Теги будут регулировать те элементы, которые заключены между ними. Например, если какой-либо текст будет заключен между приведенными в пример тегами, то он будет больше остального текста по размеру. Чаще используются два парных тега: первый открывает команду (начальный, открывающий тег), второй — закрывает (конечный, закрывающий тег). Набор и рекомендуемые интерпретации тегов определены организацией *World Wide Web Consortium, W3C* (Консорциум Всемирной паутины).

продуктом человеческого понимания текста. Например, в отношении части речи того или иного слова могут быть разные точки зрения, тогда как пол говорящего или пишущего является объективно верифицируемым.

Некоторые ученые, в частности Джон Синклер (Sinclair 1991), отрицают необходимость аннотирования вообще, считая, что следует исследовать лишь «чистый» корпус без каких-либо «примесей» посторонней информации, которая может отражать личные пристрастия и даже ошибки аннотирующего. Для других аннотирование – это возможность сделать корпус более полезным для языковых исследований. Этим можно объяснить широкое распространение таких корпусов как Британский национальный корпус и *LOB*, к достоинствам которых относится не только их содержание в виде текстов, но и аннотации языковых элементов. Тони МкЭнери (McEnery 2003) приводит следующие доводы в пользу аннотирования:

– облегчает и ускоряет поиск информации для анализа языковых данных;
– можно использовать повторно;
– корпус, аннотированный для одной цели, можно использовать для других целей;
– позволяет непосредственно фиксировать лингвистический анализ;
– создает ресурс образцов для сравнения, с помощью которых можно проводить последующие сопоставительные исследования.

Аннотирование производится тремя способами:

– автоматическая аннотация – с помощью специальных программ можно осуществлять разные виды аннотаций (частеречное таггирование, лемматизация и пр.) для быстрой и достаточно надежной обработки данных. Иногда может понадобиться ручное редактирование;

– компьютеризированная аннотация – полуавтоматический процесс с участием человека и машины дает самые надежные результаты, но требует больше времени и затрат;
– ручное аннотирование – используется, когда нет доступа к специальным программам или, когда их точность вызывает сомнения. Поскольку оно требует много времени и затрат, его целесообразно использовать только для малых корпусов.

В корпусной лингвистике в отношении аннотаций известны так называемые «максимы Лича». Их семь:

1. Должна быть возможность извлечь аннотацию из аннотированного корпуса и вернуть его в исходное (необработанное) состояние.

2. Должна быть возможность извлечь сами аннотации из текста.

3. Схема аннотации должна основываться на рекомендациях, доступных для конечного пользователя.

4. Должно быть ясно, как и кем было выполнено аннотирование.

5. Конечному пользователю должно быть известно, что аннотация корпуса не гарантирована от ошибок и неточностей и является лишь потенциально удобным инструментом.

6. Схема аннотация должна быть, насколько это возможно, основана на общепринятых и теоретически нейтральных принципах.

7. Ни одна схема аннотации на может *a priori* считаться стандартом. Стандарты возникают в ходе согласованной практики.

Аннотации могут быть не только «встроенными» в текст, но и сохраняться отдельно в виде документов SGML/XML, привязанных к оригиналу и другим аннотационным документам.

Корпусная аннотация может осуществляться на разных уровнях и в разных формах:

– на фонологическом уровне в форме границы слогов (фонетическая аннотация) обозначения интонации (просодическая аннотация);
– на морфологическом уровне в виде обозначения префиксов, суффиксов и основ (морфологическая аннотация);

– на лексическом уровне для указания на части речи (частеречное таггирование), базовые формы (лемматизация) или семантические поля (семантическая аннотация);
– на синтаксическом уровне в виде синтаксического анализа (парсинг[49], деревья зависимостей или группирование в скобки);
– на дискурсивном уровне аннотации могут указывать на анафорические отношения между частями текста, при которых в смысл одного элемента входит отсылка к другому (кореферентная аннотация), или на прагматическую информацию, такую как речевые акты (прагматическая аннотация) или стилистические особенности, например, прямая и косвенная речь (стилистическая аннотация).

На сегодняшний момент самыми проработанными и распространенными из них являются частеречное теггирование и синтаксический парсинг.

4.2.2.1 Токенизация

Любая компьютерная обработка языка начинается с сегментации текста на предложения и более мелкие единицы, называемые **токены** (*token*). Токенами могут быть слова, числа, знаки препинания и пр., т.е. символьные последовательности, отделяемые друг от друга пробелами. Процесс разделения текста на токены называется **токенизация** (*tokenization*). Ниже приведен пример токенизации предложения из корпуса журнала *Wall Street Journal*:

“Most customers don’t want to sit in a turboprop for 2 1/2 to three hours,” Mr.Lowe said.

С помощью тегов <S> и </S> для обозначения начала и конца предложения и новой строки для разграничения слов токенизация данного предложения может быть выполнена следующим образом:

[49] **парсинг** (англ. parsing) – автоматический грамматический анализ, переводящий выражения языка-объекта в выражения метаязыка описания – внутреннего языка блока анализа (Баранов 2001).

```
<S>
“
Most
customers
do
n’t
want
to
sit
in
a
turboprop
for
2½
to
three
hours
,
”
Mr.
Lowe
said
.
</S>
```

Следует отметить, однако, что такая система эффективна для алфавитных языков, где слова отделены друг от друга пробелами, но непригодна для идиографических языков (например, китайского). Токенизация текстов в алфавитных языках также вызывает ряд проблем, основными из которых являются следующие:

1) полилексемные сочетания (*multiwords*) – когда одна морфосинтаксическая единица состоит из нескольких слов, например, английское выражение *in spite of* состоит из трех слов, но обычно индексируется как один предлог, т.е. как одна морфосинтаксическая единица;
2) сокращенные формы (*mergers*) – когда одно слово соответствует нескольким морфосинтаксическим единицам, например, *can’t, gonna* и т.п.;
3) неустоявшая орфография сложных слов (*variably spelt compounds*) – в зависимости от конкрутной ситуации одному или нескольким словам соответствует одно или несколько морфосинтаксических единиц, например, *eye strain, eye-strain* или *eyestrain.*

4.2.2.2 Лемматизация

Следующим важным шагом является **лемматизация** или **лексемная аннотация** (*lemmatization / lexeme annotation / lemma annotation*). Если при грамматической аннотации такие единицы как *eat, eats, ate, eaten* и *eating* индексируются разными тегами в соответствии с разными морфосинтаксическими функциями (причастие прошедшего времени, причастие настоящего времени и т.д.), то при лексической аннотации они будут проиндексированы одним и тем же тегом, поскольку все входят в одну и ту же лемму EAT (термин «лемма» в даном случае в какой-то степени совпадает с понятием «словарная статья»). В английском языке леммная аннотация может показаться избыточной в силу его простой флективной морфологии, однако в других языках, таких как русский, имеющих широкую флективно-морфологическую систему, такая аннотация позволяет получить весьма богатую информацию для исследования всех вариантов лексемы без необходимости вручную вводить все возможные варианты для определения частотности и распределения данной лексемы.

Несмотря на наличие специальных программ леммного аннотирования – так называемых **лемматизаторов** (*lemmatizers*) – они используются далеко не во всех широко известных корпусах. В качестве примера можно привести корпус *SUSANNE*[50], содержащий лемматизированные формы слов. В следующем примере четвертый столбец содержит лемматизированые слова:

```
N12:0510g -      PPHS1m      He          he
```

[50] SUSANNE (сокр. от *Surface And Underlying Structural ANalyses of Natural English*) – подкорпус Брауновского корпуса, состоящий из 130 000 слов (64 из 500 текстов Бруновского корпуса), целью которого является установление набора стандартов аннотации и разработки достаточно подробной и эксплицитной нотации для всех аспектов поверхностной и логической грамматики реального английского языка, так чтобы разные исследователи, применяя эти стандарты к одному и тому же тексту могли получить одинаковые результаты. Аннотация проводилась вручную специалистами в области лингвистики и информатики.
Корпус свободно доступен для исследователей (http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/susanne/0.html).

```
N12:0510h –     VVDv        studied      study
N12:0510i –     AT          the          the
N12:0510j –     NN1c        problem      problem
N12:0510k –     IF          for          for
N12:0510m –     DD221       a            a
N12:0510n –     DD222       few          few
N12:0510p –     NNT2        seconds      second
N12:0520a –     CC          and          and
N12:0520b –     VVDv        thought      think
N12:0520c –     IO          of           of
N12:0520d –     AT1         a            a
N12:0520e –     NNc         means        means
N12:0520f –     IIb         by           by
N12:0520g –     DDQr        which        which
N12:0520h –     PPH1        it           it
N12:0520i –     VMd         might        may
N12:0520j –     VB0         be           be
N12:0520k –     VVNt        solved       solve
N12:0520m –     YF          +.           -
```

4.2.2.3 Частеречная аннотация

Частеречная аннотация (*part-of-speech tagging / POS tagging / tagging*), также известная как **грамматическая индексация** (*grammatical tagging*), представляет собой добавление к каждому слову корпуса соответствующей морфосинтаксической информации. Такое добавление информации эксплицирует грамматическую категорию, к которой относится каждое слово, с помощью таких кодов как: «прилагательное», «сравнительная форма»; «существительное», «исчисляемое», «единственное число»; «глагол», «настоящее неопределенное время», «3 лицо единственного числа». Индексируется и пунктуация. Разные наборы тегов могут индексировать разное число категорий и, соответственно, включать в себя разное число тегов, при этом для одних и тех же категория могут использоваться совершенно разные коды.

По справедливому замечанию Сьюзан Ханстон (Hunston 2002:82) аннотация должна осуществляться автоматически, иначе труд, затраченный на ручное добавление тегов, перевесит все преимущества. Частеречная аннотация была первым видом

аннотаций, выполненных автоматически. В 1971 году программа *TAGGIT*, разработанная в Брауновском университете, уже имела точность в 77%. В наши дни, частеречная аннотация достигла высочайшей точности – программа *CLAWS* (*Constituent Likelihood Automatic Wordtagging System*), разработанная в Ланкастерском университете, имеет коэффициент погрешности всего 4%-2%.

Ниже приведен пример частеречной аннотации из корпуса *Spoken English Corpus* с использованием набора тегов C7, содержащего 137 категорий. Пример взят из книги современной британской писательницы Джилли Купер (*Jilly Cooper*) «Поло» (*Polo*), 1991 г.:

Исходное предложение (без аннотации):

Perdita, covering the bottom of the lorries with straw to protect the ponies' feet, suddenly heard Alejandro shouting that she better dig out a pair of clean breeches and polish her boots, as she'd be playing in the match that afternoon.

Аннотированное предложение:

```
Perdita&NN1-NP0;  ,&PUN;  covering&VVG;  the&AT0;
bottom&NN1;  of&PRF;  the&AT0;  lorries&NN2;  with&PRP;
straw&NN1;  to&TO0;  protect&VVI;  the&AT0;  ponies&NN2;
'&POS;  feet&NN2;  ,&PUN;  suddenly&AV0;  heard&VVD-VVN;
Alejandro&NN1-NP0;  shouting&VVG;  that&CJT;  she&PNP;
better&AV0;  dig&VVB;  out&AVP;  a&AT0;  pair&NN0;  of&PRF;
clean&AJ0;  breeches&NN2;  and&CJC;  polish&VVB;  her&DPS;
boots&NN2;  ,&PUN;  as*CJS;  she&PNP;  'd&VM0;  be&VBI;
playing&VVG;  in&PRP;  the&AT0;  match&NN1;  that&DT0;
afternoon&NN1;  .&PUN;
```

Используемые коды:
AJ0: general adjective
AT0: article, neutral for number
AV0: general adverb
AVP: prepositional adverb
CJC: co-ordinating conjunction
CJS: subordinating conjunction
CJT: *that* conjunction
DPS: possessive determiner
DT0: singular determiner
NN0: common noun, neutral for number
NN1: singular common noun
NN2: plural common noun
NP0: proper noun

POS: genitive marker
PNP: pronoun
PRF: *of* PRP: prepostition
PUN: punctuation
TO0: infinitive *to*
VBI: *be*
VM0: modal auxiliary
VVB: base form of lexical verb
VVD: past tense form of lexical verb
VVG: -ing form of lexical verb
VVI: infinitive form of lexical verb
VVN: past participle form of lexical verb

Все теги содержат три символа и прикреплены к словам согласно стандарту *TEI* с помощью ограничительных служебных символов `&` и `;`. Интересно, что некоторые слова, например, *heard*, имеют два тега – `VVD` и `VVN`. Это так называемые «складные теги» или «теги-бумажники» (*portmanteau tags*), используемые для того, чтобы помочь пользователю в ситуациях, когда есть большая вероятность того, что компьютер может выбрать неверную часть речи из имеющегося набора (этот корпус не был откорретирован вручную).

4.2.2.4 Семантическая разметка

Еще большую трудность представляет семантический анализ корпуса. Тони МакЭнери и Эндрю Уилсон (McEnery, Wilson 2001) выделяют два типа семантической аннотации: разметка семантических связей между объеками текста, например, агенса и патиенса какого-либо действия (такая аннотация используется достаточно редко), и разметка в той или иной форме семантических признаков слов текста, в частности, значений слова. Такой подход имеет более широкое распространение и уходит корнями в 1960-е годы, когда за основу брался знаменитый тезаурус Роже (*Roget's Thesarus*), в котором слова были организованы по широким семантическим категориям.

Теги, используемые в семантической аннотации, указывают на семантические поля, объединяющие слова в группы на основе того, насколько они связаны с конкретным ментальным концептом. В эти группы входят не только синонимы и антонимы, но также гиперонимы и гипонимы (Bianchi 2012). Более того, как и в отношениях между гиперонимами и гипонимами, возможны разные «уровни» абстракции, например, слово «кошка» при необходимости может быть помечено, как «семейство кошачих», «млекопитающее», «животное» и даже «живое существо». Это называется «гранулярность» (*granularity*) или «степень детализации» (*delicacy of detail*), при этом выбор уровня гранулярности определяется не столько теоретическими, сколько прагматическими факторами (Wilson, Thomas 1997).

Основной сложностью семантической аннотации является проблема многозначности. Значение слова легко определяется человеком на основе контекста или ситуации, но является трудной задачей для машины. Начиная с 1950 гг., специалисты в области машинного перевода и обработки естественного языка (*natural language processing, NLP*), работают над алгоритмами систем разрешения лексической многозначности (*word sense disambiguation, WSD*).

Обычно система *WSD* выбирает из набора всех возможных значений слова то значение, которое соответствует данному контексту. Например, она распознает слово *bank* как «финансовое учреждение», если обнаружит, что в окружающем слово тексте говорится о финансовых вопросах, и как «берег реки», если речь идет о реке. Некоторые системы могут даже разграничить «банк» как учреждение от «банка» как здание, в котором расположено это учреждение. (Raysona et al. 2004). Проблема лексической многозначности обычно решается с помощью нескольких приемов: частеречная разметка, которая обычно предшествует семантической аннотации; статистическая информация в отношении частотности и контекстно-зависимые правила. И, наконец, если слово попадает в несколько семантических полей, ему присваивается несколько

индексов, а затем, на основе частотности или рассмотрения проблемной области, выбирается наиболее оптимальный.

Несмотря на то, что ученые единодушно признают, что «идеальная» система семантической аннотации не существует, ими (см., в частности, Wilson, Thomas 1997) выдвинут ряд критериев для создания или выбора оптимальной системы семантической аннотации:

1) она должна быть обоснованной и с точки зрения общей лингвистики, и с точки зрения психолингвистики;

2) она должна в полной мере отражать словарный состав всего корпуса, а не только его части;

3) она должна быть достаточно гибкой для внесения корректировок, связанных с изменением рассматриваемого временного периода, языка, регистра или текстовой базы;

4) она должна действовать на соответствующем уровне гранулярности (степени детализации);

5) она должна, при необходимости, иметь иерархическую структуру;

6) при наличии стандартов она должна им соответствовать.

Рассмотрим одну из таких систем, разработанную исследовательским центром *UCREL*[51] (*University Centre for Computer Corpus Research on Language*), действующим на базе кафедры лингвистики и современного английского языка (*Department of Linguistics and Modern English Language*) и кафедры кафедры информационно-вычислительных систем (*Department of Computing*) Ланкастерского университета. Система, разработанная в этом центре, также называется *UCREL*. В 1988 году центр занялся разработкой

[51] *UCREL* возник в 1970 г., когда Джеффри Лич возглавил группу под названием *CAMET (Computer Archive of Modern English Texts)* в рамках кафедры английского языка Ланкастерского университета. Задачей группы было создание электронного корпуса британского варианта английского языка объемом в 1 млн слов параллельно Брауновскому корпусу американского варианта. В проекте, завершившемся в 1978 году, приняли участие норвежские университеты Осло и Бергена, и поэтому новый корпус получил название *Lancaster/Oslo-Bergen* (сокращенно *LOB*).

программ семантического анализа (разметкой слов по семантическим полям). Проекту под названием ACASD (*Automatic content analysis of spoken discourse*) удалось добиться точности семантической разметки выше 90%.

В 1990 году центр начал еще один проект семантического аннотирования – *USAS (UCREL semantic analysis system*), изначальный набор тегов которого основывался на лексиконе *Longman Lexicon of Contemporary English* Тома МкАртура (McArthur 1981), предлагавшем наиболее оптимальную классификацию тезаурусного типа с точки зрения практических задач разметки, возникших в ходе исследования. Окончательный переработанный набор тегов был упорядочен иерархически по 21 основной дискурсивной области (см. таблицу 4), каждая из которых состояла из подобластей более низкого уровня (всего 232 категории).

Таблица 4. Семантическая классификация в системе USAS.

A General and abstract terms	**B** The body and the individual	**C** Arts and crafts	**E** Emotion
F Food and farming	**G** Government and the public domain	**H** Architecture, buildings, houses and the home	**I** Money and commerce in industry
K Entertainment, sports and games	**L** Life and living things	**M** Movement, location, travel and transport	**N** Numbers and measurement
O Substances, materials, objects and equipment	**P** Education	**Q** Linguistic actions, states and processes	**S** Social actions, states and processes
T Time	**W** The world and our environment	**X** Psychological actions, states and processes	**Y** Science and technology
Z Names and grammatical words			

Далее лексемы из каждого семантического поля делятся по областям значения, отражающих отношения синонимии/антонимии, гиперонимии/гипонимии или меронимии/холонимии[52]. Например, поле {F: Food and Farming} далее разделяется на четыре более узких области:

F: FOOD & FARMING

F1 ***Food****:* Термины, относящиеся к еде и приготовлению пищи

Примеры*: afters, bacon, banana, befores, breakfast, butter, casseroled, cereal, chilli, cook, afternoon tea, apple sauce, after dinner mint, canteen meal, chewing gum, cooking facilities, dairy product*

F2 ***Drinks****:* Термины, относящиеся к напиткам и употреблению напитков

Примеры: *alcoholic, ale, beer, beverage, boozing, cola, coffee, cuppa, inebriated (++), temperance (-), apple juice, cherry coke, cup of coffee, drinking chocolate, glass of wine, hit the bottle, liqueur coffee, mineral water, on the wagon (-), pub crawl, Tia Maria, tonic water*

F3 ***Cigarettes and drugs****:* Термины, относящиеся к сигаретам и наркотикам, включая их воздействие

Примеры: *cannabis, cigar, detox, drugged, e-ing, LSD, non-addictive, OD,tobacco, pipe, heroin, cocktail cigarette, drug addiction, glue sniffing, hard drug, non smoking, passive smoking, take a puff*

F4 ***Farming & Horticulture*** Термины, относящиеся к сельскому хозяйству и садоводству

Примеры: *agricultural, beehive, compost, dairy, farming, forestry, gardening, harvest, bee keeping, estate management, free range, grounds maintenance, landscape gardening, stud farm*

[52] Мерономия – это классификация явлений, основанная на отношениях меронимии и холонимии. **Мероним** – понятие, которое является составной частью другого (другое название – партонимом (*англ*. part = «часть»); **холоним** – понятие, которое является целым над другим(и) понятием(ями) (то есть другое(ие) понятие(я) предстает(ют) в качестве составной части первого). Например, термины *двигатель, колесо и капот* представляют собой меронимы по отношению к термину *автомобиль*. В свою очередь, термин *автомобиль* выступает холонимом по отношению к терминам *двигатель, колесо и капот*. (Кронгауз 2005).

Как видим, слова *bacon, cereal, beer, coffee, cigar, tobacco, beehive, heroin* входят в одно семантическое поле {Food and Farming}, но далее распределяются по различным подобластям значения: *Food, Drinks, Cigarettes and drugs, Food and Farming and Horticulture*.

Система *USAS* состоит из трех модулей:

1. *CLAWS* – частеречная аннотация (каждому слову или словосочетанию присваивается тег части речи).

2. *SEMTAG* – семантическая аннотация (каждому слову или словосочетанию присваивается тег (или несколько тегов через косую черту, если у них несколько значений) семантической категорией).

3. *AUXRULE* – подмодуль *SEMTAG* (разрешение многозначности смысловых и вспомогательных глаголов *be, do* и *have* на базе непосредственной коллокации (или отсутствия коллокации) с соответствующими причастными формами).

Следующий пример иллюстрирует использование семантической аннотации:

```
PPIS1     I           Z8
VV0       like        E2+
AT1       a           Z5
JJ        particular  A4.2+
NN1       shade       O4.3
IO        of          Z5
NN1       lipstick    B4
```

Текст читается сверху вниз. Грамматическая аннотация дана слева, а семантическая – справа. Семантические теги содержат следующее:

– прописная буква, указывающая на общее семантическое поле;

– цифра, указывающая на первую подобласть поля;

– (факультативно) знак десятичной дроби, за которым следует цифра, указывающая на еще более узкую подобласть;

– (факультативно) – один или несколько знаков «плюс» или «минус», указывающие на положительное или отрицательное положение на семантической шкале.

Например, **A4.2+** показывает, что слово входит в категорию «общие и абстрактные слова» (**A**), подкатегорию «классификация» (**A4**), под-подкатегорию «конкретное и общее» (**A4.2**) и в группу «конкретное» в противоположность «общему» (**A4.2+**). Аналогичным образом **E2+** принадлежит категории «эмоциональные состояния, действия, события и процессы» (**E**), подкатегории «симпатия и антипатия» (**E2**) и относится к «симпатии», в отличие от «антипатии» (**E2+**).

Семантическая аннотация предназначена для полнозначных слов «открытого класса» (т.е. пополняемого новыми словами). Слова, принадлежащие к закрытым классам, а также имена собственные, обозначаются тегом, начинающимся с буквы Z, и не включаются в статистический анализ.

4.2.2.5 Синтаксический анализ (парсинг)

Частеречная аннотация, о которой говорилось выше, служит основной не только семантической аннотации, но и для синтаксического анализа. После разметки морфосинтаксических категорий их можно исследовать на более высоком уровне синтаксических отношений (McEnery, Wilson 2001), то есть анализировать предложения корпуса с точки зрения их составляющих. Такая процедура получила названия парсинг (*parsing*). В отличие от более низких уровней аннотации, он способен выявлять структурные и даже функциональные аспекты языка.

Неудивительно, что синтаксически размеченные корпусы обычно меньше по объему, чем обычные, поскольку большинство из них проходит ручную проверку точности. Первым синтаксически размеченным корпусом был *Lancaster-Leeds Treebank* – проанализированный вручную небольшой (45000 слов) подкорпус корпуса *LOB*, послуживший для обучения компьютерных программ,

которые позже использовались для создания корпуса *Lancaster Parsed Corpus* (*LPC*), другого подкорпуса *LOB*, объем которого составил уже порядка 144 000 слов. В разных вариантах корпуса *Penn Treebank* (содержавшего новостные сообщения, тексты Брауновского корпуса, учебную литературу, записи радиопередач и затранскрибированные тексты корпуса *Switchboard Corpus*) использовали различные виды словосочетаний и клауз, отражавших ланкастерский подход.

Корпус Джефа Сэмпсона *SUSANNE* (*Surface and Underlying Structural Analyses of Naturalistic English*), состоящий из 130 000 слов, представляет собой бесплатно скачиваемый американский подкорпус Брауновского корпуса, который имеет теги не только для словосочетаний и клауз, но и теги поверхностных функций для для таких ролей как «логическое подлежащее» или «агенс пассивной консутркции». Особенностью другого корпуса Сэмпсона – *LUCY* (165,000 words), состоящего, помимо прочего, из текстов экзаменов по английскому языку, курсовых работ, сочинений и детской письменной речи, адаптирован к работе с нестандартными структурами, в отличие от «правильных» отредактированных текстов, для работы с которыми были обучены машинные парсеры. Еще один корпус Сэмпсона – *CHRISTINE* (80500 слов), основанный на подкорпусе *BNC, London-Lund Corpus* и корпусе *Reading Emotional Speech Corpus* – отличается тем, что имеет дело с особенностями устной речи, такими как паузы, дискурсивные маркеры, исправление собственных ошибок в речи и т.д. Корпус *ICE-GB* – британский компонент корпуса *International Corpus of English* (полностью синтаксически размеченый и выверенный вручную) является прекрасным ресурсом для исследователей устного и письменного английского языка.

Синтаксический анализ может варьироваться от небольших блоков (*chunks*), подобных словосочетаниям, до полного иерархического парсинга, включая функциональный анализ.

В зависмости от разных моделей грамматического анализа, взятых за основу, основными задачами парсинга являются следующие:

– четкое определение слов, используемых в предложении;
– присвоение словам соответствующего синтаксического описания;
– определение границ групп и клауз;
– распределение групп по компонентам клауз;
– группированиие групп и клауз для определения синтаксических составляющих предложения;
– соответствующее обозначение составляющих.

Большинство систем синтаксического анализа, разработанных для английского языка, ориентированы на те или иные принципы формальной лингвистики, такие как теория управления и связывания (Government and Binding theory), контекстно-свободная грамматика (Context-Free Grammar) Н. Хомского, древоприсоединяющая грамматика (Tree-adjoining grammars) А. Джоши и др., для выявления внутренних синтаксических отношений, лежащих в основе составляющих, создающих предложение.

Стандартным способом представления синтаксической структуры грамматического предложения является «синтаксическое дерево» или «дерево зависимостей» (*syntax tree / parse tree*), отражающее все шаги образования предложения из корневого узла. Это означает, что каждый внутренний узел дерева представляет применение какого-либо грамматического правила. Поэтому синтаксически размеченные корпусы иногда называют «банк синтаксических деревьев» (*treebank*). Например, предложение из корпуса *BNC* "Claudia sat on a stool." можно представить в виде следующей древовидной диаграммы:

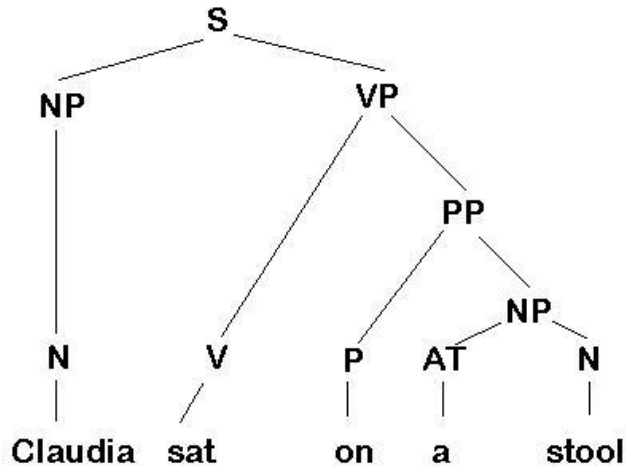


Рис. 2. Образец «синтаксического дерева»

Где S – предложение, NP – именная группа, VP – глагольная группа, PP – предложная группа, N – существительное, V – глагол, AT – артикль, P – предлог.

Однако такие объемные диаграммы редко встречаются в корпусных аннотациях, чаще всего аналогичная информация выражается скобочными записями с метками. Например, приведенное выше предложение будет выглядеть следующим образом:

```
[S[NP Claudia_NP1 NP][VP sat_VVD [PP on_II [NP a_AT1 stool_NN1 NP] PP] VP] S]
```

Морфосинтаксическая информация присоединяется к словам с помощью знаков с нижним подчеркиванием (_) в форме частеречных тегов, в то время как составляющие обозначаются открывающими и закрывающими квадратными скобками с указанием типа группы в начале и в конце, напр., [S ..... S].

Иногда такие скобочные аннотации изображаются с помощью абзацных отступов, что несколько напоминает древовидную диаграмму (эта система используется, в частности, в проекте *Penn Treebank*), например:

```
[S
     [NP Claudia NP]
     [VP sat
            [PP on
                  [NP a stool NP]
            PP]
      VP]
S]
```

Существуют разные системы грамматического анализа, отличающиеся между собой двумя основными факторами:

– количество типов составляющих, используемых системой;

– каким образом типы составляющих могут объединяться друг с другом.

Однако, несмотря на эти различия, большинство систем синтаксического анализа основано на той или иной форме контекстно-свободной фразовоструктурной грамматики. При этом основное различие проходит между полным синтаксическим анализом (*full parsing*) и минимальным, так называемым «скелетным» синаксическим анализом[53] (*skeleton parsing*).

Полный парсинг нацелен на максимально детальный анализ структуры предложения, в то время как скелетный парсинг носит менее подробный характер, что проявляется в использовании менее детального набора типов синтаксических составляющих и игнорировании, к примеру, внутренней структуры определенных типов составляющих. Следующие примеры иллюстрируют различия между ними.

[53] Данный упрощенный грамматический анализ используется в данное время в проекте *UCREL*. Тексты анализируются вручную с помощью программы *EPICS*, созданной Роджером Гарсайдом. Эта программа ускоряет ручную обработку за счет того, что в ней хранится набор составляющих, которые открываются в определенной точке текста, а затем оператор (человек) закрывает эти составляющие или открывает дополнительные в соотвествующих точках. Задача *EPICS* – обеспечить парсинг с минимальным количеством нажатия на клавиши: при полном объеме оператор может разметить предложение длиной более 20 слов менее чем за одну минуту (Garside, Leech 1991).

Исходное предложение (из корпуса *Lancaster-Leeds treebank)*:

Another new style feature is the wine-glass or flared heal, which was shown teamed up with pointed, squared, and chisel toes.

Полный парсинг (McEnery, Wilson 1996):

```
[S[Ncs another_DT new_JJ style_NN feature_NN Ncs] [Vzb
is_BEZ Vzb] [Ns the_AT1 [NN/JJ& wine-glass_NN [JJ+ or_CC
flared_JJ HH+]NN/JJ&] heel_NN ,_, [Fr[Nq which_WDT Nq] [Vzp
was_BEDZ shown_VBN Vzp] [Tn[Vn teamed_VBN Vn] [R up_RP R] [P
with_INW [NP[JJ/JJ/NN& pointed_JJ ,_, [JJ- squared_JJ JJ-] ,_,
[NN+ and_CC chisel_NN NN+]JJ/JJ/NN&] toes_NNS Np]P]Tn]Fr]Ns]
._. S]
```

Структура синтаксических составляющих здесь показана через двойные помеченные квадратные скобки, а слова имеют частеречные метки. Синтаксические составляющие обозначены следующим образом:

& сочинительная конструкция в целом
+ подчинительный конъюнкт присутствует
– подчинительный конъюнкт отсутствует
Fr группа придаточного относительного
JJ группа прилагательного
Ncs группа существительного, исчисляемое существительное в ед. числе
Np группа существительного, исчисляемое существительное во мн. числе
Nq группа существительного, относительное местоимение
Ns группа существительное, единственное число
P предложная группа
R адвербиальная группа
S предложение
Tn группа причастия прошедшего времени
Vn глаголная группа, причастие прошедшего времени
Vzb глагольная фраза, третье лицо ед. число глагола *to be*
Vzp глагольная фраза, страдательный залог 3-е лицо ед. число

Исходное предложение (из корпуса *Spoken English Corpus)*:

For the members of this university this charter enshrines a victorious principle; and the fruits of that victory can immediately be seen in the international community of scholars that graduated here today.

Скелетный парсинг:

```
S& [P For_IF [N the_AT members_NN2 [P of_IO [N this_DD1
university_NNL1 N]P]N]P] [N this_DD1 charter_NN1 N] [V
enshrines_VVZ [N a_AT1 victorious_JJ principle_NN1 N]V]S&] ;_;
and_CC [S+[N the_AT fruits_NN2 [P of_IO [N that_DD1
victory_NN1 N]P]N] [V can_VM immediately_RR be_VB0 seen_VVN
[P in_II [N the_AT international_JJ community_NNJ [P of_IO [N
scholars_NN2 N]P] [Fr that_CST [V has_VHZ graduated_VVN
here_RL today_RT V]Fr]N]P]V]S+] ._.
```

Оба примера в целом схожи, но во втором именные группы просто обозначены буквой *N*, тогда как в полном парсинге показаны несколько типов именных групп, различающихся между собой по ряду признаков, таких как число. В скелетном парсинге представлены только следующие обозначения составляющих:

Fr относительная клауза
N именная группа
P предложная группа
S& 1-й главный конъюнкт сложносочиненного предложения
S+ 2-й главный конъюнк сложносочиненного предложения
V глагольная группа

Тем не менее, не все корпусы проходят синтаксический анализ на основе контекстно-свободной фразовоструктурной грамматики. Например, при частеречной аннотации и парсинге корпуса *Birmingham Bank of English* использовалась форма грамматики зависимостей, известной как грамматика ограничений (Karlsson et al. 1995). Вместо того, чтобы определять иерархии типов составляющих групп, грамматика ограничений выделяет грамматические функции слов внутри предложения и взаимозависимости между ними. Например, код со стрелкой вправо (AN>) показывает левое определение, в данном случае выраженное прилагательным, тогда как

код со стрелкой влево (<NOM-OF) показывает правое определение, в данном случае с предлогом *of*. Ниже приведен пример применения парсера *Constraint Grammar Parser of English* к корпусу *Helsinki corpus*:

Исходное предложение:

It has maintained its independance and present boundaries intact since 1815.

Парсинг грамматики ограничений:

```
"<it>"
   "it" <*> <NonMod> PRON NOM SG3 SUBJ @SUBJ
"<has>"
   "have" <SVO> <SVOC/A> V PRES SG3 VFIN @+FAUXV
"<maintained>"
   "maintain" <Vcog> <SVO> <SCOC/A> PCP2 @-FMAINV
"<its>"
   "it" PRON GEN SG3 @GN>
"<independence>"
   "independence" <-Indef> N NOM SG @OBJ @NN>
"<and>"
   "and" CC @CC
"<present>"
   "present" <SVO> <P/in> <P/with> V INF @-FMAINV
   "present" A ABS @AN
"<boundaries>"
   "boundary" N NOM PL @OBJ
"<intact>"
   "intact" A ABS @PCOMPL-O @<NOM
"<since>"
   "since" PREP @<NOM @ADLV
"<1815>"
   "1815" <1900> NUM CARD @<P
<$.>"
```

Рядом с каждом словом даны три (или больше) единицы информации. Первый элемент в двойных кавычках это лемма слова, за которой идет код части речи (который может включать несколько

последовательностей (напр. N NOM PL), а в правой стороне строки имеется код грамматической функции слова. Он начинается с @ и означает следующее:

@+FMAINV	главный предикатор, личная форма
@-FMAINV	главный предикатор, неличная форма
@	прилагательное-премодификатор
@CC	сочинительной союз
@DN>	детерминатив
@GN>	генетив-премодификатор
@INFMARK>	маркер инфинитива
@NN>	существительное-премодификатор
@OBJ	дополнение
@PCOMPL-O	объектное дополнение
@PCOMPL-S	субъектное дополнение
@QN>	левый квантификатор
@SUBJ	подлежащее

Таким образом, если разметка содержит контекстную (экстралингвистическую) информацию из формальных и объективных источников и носит унифицированный характер, то лингвистическая аннотация требует тщательного лингвистического анализа специалистом и потому, в силу субъективности интерпретации, может быть неодинаковой и разноречивой.

Особая проблема возникает при подготовке корпусов параллельных текстов. Она заключается в установлении соответствий между текстом оригинала и его переводами. Для решения такой задачи используется так называемый метод автоматического **выравнивания** (*alignment*) текстов. Его суть заключается в параллельной сегментации оригинального текста и его перевода по предложениям.

§ 4.3. Требования к корпусу

Если рассматривать отдельные тексты, из которых будет создан корпус, как данные – необработанные материалы и факты, то результатом их обработки будет полученная информация, которая должна отвечать определенным требованиям. Специалисты в области информационных технологий (Всеволодова 2007) к числу таких требований относят:

1. Адекватность информации – степень соответствия информации, полученной потребителем, тому, что автор вложил в ее содержание (т.е. в данные). При этом, в силу того, что информация является продуктом взаимодействия данных и методов, то на ее свойства влияют как адекватность данных, так и адекватность методов.

2. Полнота информации – достаточность информации для принятия решения. Она зависит как от полноты данных, так и от наличия необходимых методов.

3. Избыточность информации – превышение количества информации, используемой для передачи или хранения сообщения, над его информационной энтропией (количеством информации на символ передаваемого сообщения). Наличие избыточности способствует повышению помехоустойчивости сообщений. Более того, избыточность информации позволяет повышать ее достоверность за счет применения специальных методов, в том числе и основанных на теории вероятностей и математической статистике: в результате отсева объем данных сокращается, но их достоверность увеличивается.

4. Объективность и субъективность информации – в силу того, что методы всегда являются субъективными, объективность информации можно назвать относительной. Более объективной принято считать ту информацию, в которую методы вносят меньший субъективный элемент.

5. Доступность информации – возможность получить ту или иную информацию. На степень доступности влияют одновременно

доступность данных и доступность адекватных методов для их интерпретации. При отсутствии адекватных методов могут применяться неадекватные методы, ведущие к неполной, неадекватной или недостоверной информации.

6. Актуальность информации – степень соответствия информации текущему моменту времени.

В соответствии с этим, корпус, служащий целям исследования и анализа языка, также должен удовлетворять определенным критериями, к числу которых разные ученые (Aarts 1991, Biber et al. 1998, Knowles 1996, McEnery & Wilson 1996, Sinclair 1991, 1996; Tognini-Bonelli 2001) относят следующие:

1. **Аутентичность** (*authenticity*) – для того, чтобы отражать реальное функционирование языка, тексты должны быть получены из реальной комуникации[54]. Несоблюдение этого требования, в свое время, привело к созданию традиционных прескриптивных грамматик. Аутентичный текст – это текст, возникший естественным образом, а не созданный для того, чтобы продемонстрировать то или иное языковое явление (Bowker, Pearson 2002; Buendgens-Kosten 2014; Widdowson 1979).

2. **Репрезентативность (представительность)** (*representativeness*) – в общем виде это способность корпуса достоверно представлять ту или иную разновидность языка. Рэнди Риппен (Rippen 2010) формулирует это в виде вопроса: Собрал ли я достаточно текстов (слов), чтобы точно представлять исследуемый тип языка? В некоторых случаях есть возможность полного представления изучаемой

[54] Джон Синклер (Sinclair 1996) формулирует это следующим образом: «весь материал должен быть собран из подлинной коммуникации людей, происходящей в связи с их обычными делами, в отличие от экспериментальных или искусственных условий какого-либо рода. Все, что требует вмешательства лингвиста (за исключением минимального участия, требуемого для получения данных), является основанием для объявления корпуса «специальным». Это защищает интересы тех, кто собирается делать суждения о том, как язык используется в повседневном общении, и кто может быть введен в заблуждение, используя данные, полученные в экспериментальных условиях или искусственных обстоятельствах любого рода».

разновидности языка, например, можно собрать все произведения того или иного автора, исторические тексты определенного периода, тексты, отражающие конкретное событие (через радио и телепередачи, политические выступления и т.п.), – в этом случае достигается полная репрезентативность. Однако в большинстве случаев это невозможно, и тогда требуется определить объем корпуса, чтобы он стал достаточно представительным.

Джеффри Лич (Leech 1991) утверждает, что корпус является представительным, если результаты, полученные на основе его содержания, могут быть обобщены до более крупного гипотетического корпуса. Если, скажем, корпус должен представлять определенную разновидность языка, его можно будет считать представительным, если его результаты могут быть обобщены до этой разновидности.

Джон Синклер (Sinclair 2005) выделяет шесть этапов на пути достижения представительности корпуса:

1) выбор структурных критериев, используемых для построения корпуса, для создания на их основе структуры основных компонентов корпуса;

2) создание исчерпывающего инвентаря типа текстов для каждого компонента;

3) распределение типов текста в порядке приоритета с учетом всех факторов, которые могут увеличить или уменьшить важность типа текстов;

4) определение целевого объема каждого типа текста с учетом: а) общего целевого объема для каждого компонента, б) количества типов текстов, в) важности каждого их них и г) практической целесообразности сбора их в таком количестве;

5) когда корпус примет определенные очертания, необходимо постоянно сопоставлять фактический объем материала с первоначальным планом;

6) (самое важное) документирование этих этапов, чтобы пользователи имели образец для сравнения, если получат

неожиданные результаты, и чтобы можно было использовать имеющийся опыт для усовершенствания корпуса.

Дуглас Байбер (Biber 1993) определяет репрезентативность как то, в какой степени выборка включает весь диапазон изменчивости совокупности. При этом он выделяет выделяет два типа изменчивости – ситуативную (сюда входят разные регистры и жанры целевой совокупности, т.е. – типов текста или речевых ситуаций, которые должны будут войти в корпус) и лингвистическую (различные языковые дистрибуции, обнаруженные в совокупности). При этом он также утверждает (Biber 1990), что языковая репрезентативность зависит, прежде всего, от ситуационной, а также от количества слов на текст и от количества текстов на регистр или жанр, входящие в корпус. Проведя несколько статистических анализов, он выяснил, что наиболее распространенные языковые явления (напр., личные местоимения, сокращенные формы, предлоги, времена настоящего и прошедшего времени) встречаются с достаточно стабильной частотностью в текстах из 1000 слов. В отношении количества текстов, необходимых для адекватного представления какого-либо регистра или жанра в корпусе, он обнаружил достаточно стабильную лингвистическую тенденцию для десяти (а иногда даже пяти) текстов на жанр или регистр.

Однако необходимо иметь в виду, что репрезентативность не всегда поддается объективной оценке (Tognini Bonelli 2001). Убедиться в том, что корпус не является достаточно представительным, чаще всего можно лишь тогда, когда его результаты окажутся искаженными.

3. **Сбалансированность** (*balance*) – сбалансированный корпус, как правило, охватывает широкий диапазон текстовых категорий, которые можно считать представительными для исследуемого языка или языковой разновидности. Эти текстовые категории отбираются пропорционально для включения в корпус так, чтобы он представлял управляемо малую модель лингвистического материала, которые собираются изучать составители корпуса (Atkins et al 1992).

Большинство первых корпусов имело перевес в сторону письменных текстов или полностью состояли из них. Даже в первом мегакорпусе второго поколения *British National Corpus* только 10% из 100 миллионов слов были из устных источников. В меньшем по объему корпусе *ICE* баланс составлял 60% устных текстов и 40% письменных, сделав его одним из немногих корпусов, имевших баланс в пользу устных текстов (Kennedy 1998). Вопрос о сбалансированности возникает не только в отношении корпусов, предназначенных представлять язык в целом, но одну конкретную область, тему, жанр, и могут полностью игнорироваться корпусами, состоящими из всего, опубликованного за определенный период времени, полного собрания работ автора или другой полной совокупности текстов.

Баланс в корпусе достигается не за счет равной представленности текстов из разных источников, таких как письменные и устные. Никто не знает сооотношение устных и письменных слов, произведенных в одном конкретной языке в один конкретный день. Для каждого из нас устная речь имеет преобладание над письменной с точки зрения того, что мы продуцируем или воспринимаем ежедневно. Однако письменный текст (например, газетную статью) могут прочитать 10 миллионов человек, однако устный диалог (например, по поводу покупки товара в магизине) будет услышан только двумя людьми, принимавшими участие в разговоре.

Внутри письменных корпусов баланс также трудно достижим. Джон Синклер (Sinclair 1991) предположил, что для универсального письменного корпуса минимальным критерием отбора материала может служить деление на художественные и нехудожественные тексты; книги, журналы или газеты; официальные и неофициальные тексты; учет пола, возраста, происхождения авторов. Как достичь баланса между популярными авторами / ораторами и большинством малоизвестных? Для решения этих задач создатели корпусов разработали очень сложные механизмы (напр., для *British National Corpus*) обеспечения репрезентативности и сбалансированности.

4. **Объем** (*size*) – вопрос, каков должен быть оптимальный объем корпуса являтся одним из ключевых при проектировании корпуса[55]. Для краткости его можно свести к двум полюсам – корпус как можно большего объема (например, для лексикографических проектов) или небольшие специализированные корпусы (например, для учебных целей). Таким образом, главным фактором в определении объема корпуса должна быть его цель (Nelson 2010).

Последние десятилетия видели создание универсальных корпусов – *British National Corpus* (первое поколение) и *Bank of English* (второе поколение). Ставя перед собой цель представлять язык в целом, они были созданы так, чтобы включать самое широкое разнообразие текстов и типов текстов, как письменных, так и устных, что предполагало максимальный объем. С другой стороны, высокочастотные явления, такие как функциональные и вспомогательные слова, можно легко извлечь из статистически незначительного количества единиц небольшого корпуса. В частности, Дуглас Байбер (Biber 1993) рассчитал минимальное количество текстов, необходимых для репрезентации отдельных языковых явлений в корпусе.

Таблица 5. Расчет требуемого размера выборки (количества текстов) для корпуса в целом.

	Средняя оценка в пилотном корпусе	**Стандарт-ное откло-нение в пилотном корпусе**	**Допусти-мая погреш-ность**	**Необходи-мое коли-чество *N***
Существительные	180,5	35,6	9,03	59,8
Предлоги	110,5	25,4	5,53	81,2
Настоящее время	77,7	34,3	3,89	299,4
Прошедшее время	40,1	30,4	2,01	883,1
Страдательный залог	9,6	6,6	0,48	726,3
Придаточные определительные	3,5	1,9	0,18	452,8
Придаточные условия	2,5	2,2	0,13	1190,0

[55] Большинство корпусов имеют **конечный объем** (*finite size*), свыше которого они не увеличиваются. Однако, для динамических корпусов (monitor corpus) это требование не соблюдается и их объем постоянно пополняется, что вызывает определенные сложности, связанные с тем, что при постоянном изменении корпуса качественные исследования, сделанные в разное время, не поддаются сопоставлению.

Его расчеты были произведены на основе корпуса, состоящего из 481 текста из 23 устных и письменных регистров. Вычисления учитывали среднюю оценку, стандартное отклонение и допустимую погрешность.

Как отмечает Джеффри Лич (Leech 1991), размер корпуса не имеет первостепенного значения. Во-первых, как говорилось выше, небольшие корпусы могут содержать достаточное число примеров. Во-вторых, корпусы, для которых требуется большая ручная аннотация (напр., прагматическая аннотация), не могут быть большими из-за высокой трудоемкости, которая потребуется на их обработку. В-третьих, некоторые программы обработки корпусов устанавливают предел числа конкордансов, которые могут быть извлечены из корпуса.

4. **Формирование выборки / отбор материала** (*sampling*) – поскольку язык бесконечен, а объем корпуса конечен, при его создании неизбежно встает вопрос отбора материала. Мы не можем дать исчерпывающее описание естественного языка, поэтому для обеспечения репрезентативности и сбалансированности, требуемых для нашей исследовательской задачи, необходимо сформировать выборку[56]. Как отмечалось, корпус представляет собой выборку более крупной совокупности (*population*). Выборка считается репрезентативной, если то, что мы обнаружили, справедливо для общей совокупности данных. С точки зрения статистики, выборка представляет собой уменьшенный вариант большей совокупности. Это ставит перед исследователем задачу формирования выборки, которая, учитывала бы границы объема и как можно точно представляла бы характеристики совокупности, особенно те, которые важны для исследования. С этой целью необходимо определить единицу выборки (*sample unit*) и границы совокупности. Например, для письменного текста единицей выборки может быть книга, журнал, газета. Совокупность – это набор всех единиц выборки, а перечень единиц выборки представляет собой основу выборки (*sample frame*).

[56] См. также Wynne 2005.

Совокупностью, из которой брались образцы для первого Брауновского корпуса, были все письменные тексты, опубликованные в США в 1961 году, при этом основой выборки было собрание книг и журналов библиотеки Брауновского университета и библиотеки Провиденс Атенеум. Для корпуса *LOB* целевой совокупностью были все письменные тексты на английском языке, опубликованные в Соединенной королевстве Великобритании и Северной Ирландии в 1961 году, а основой выборки послужил Британский национальный кумулятивный библиографическай предметный индекс за 1960-1964 гг. для книг и Справочник по печати Уиллинга за 1961 г. для журналов.

При проектировании корпуса совокупность может определяться с точки зрения речепорождения, речевосприятия и речевого продукта. Первые два типа организованы в основном по демографическому признаку (возраст, пол, социальное положение говорящих / пишущих и слушающих / читающих), тогда как последний строится вокруг текстовых категорий / жанра языковых данных.

При создании Брауновского корпуса и корпуса *LOB* использовался критерий продукта речи, а в Британском национальном корпусе *(BNC*) совокупность определялась преимущественно на основе речепорождения и речевосприятия. Следует отметить чрезвычайную сложность определения совокупности или создания основы выборки по этому критерию, особенно для устной речи, не имеющей каталогов и библиографий.

После определения целевой совокупности и основы выборки можно использовать различные методы, чтобы подобрать ту выборку, которая будет как можно лучше представлять совокупность. Самый удобный метод – это простая случайная выборка (*simple random sampling*), когда все единицы выборки внутри основы нумеруются и с помощью таблицы случайных чисел происходит отбор выборки. Поскольку вероятность выбора единицы находится в прямой зависимости от ее частотности в совокупности, такой метод генерирует выборку, не включающую в себя относительно редкие единицы совокупности, даже если они интересны для исследователя.

Одним из способов преодоления этой проблемы является стратифицированная случайная выборка, при которой сначала вся совокупность делится на относительно однородные группы (страты), а затем к каждой страте применяется простой случайный выбор. В Брауновском корпусе и корпусе *LOB* целевая совокупность сначала разделялась на 15 текстовых категорий, таких как репортаж, академический текст, различные виды художественной литературы. Затем из каждой категории извлекались образцы. Демографическая выборка, когда единицы совокупности сначала группируются на основе возраста, пола, социального положения говорящего / пишущего, также является примером стратифицированной выборки. Д. Байбер (Biber 1993) отмечает, что стратифицированная выборка никогда не бывает менее репрезентативной, чем простая случайная выборка.

Следующее решение, которое требуется принять в отношении выборки, – это ее размер. Например, для письменной речи, нужно ли брать полные тексты (документы в целом) или фрагменты текста? Если фрагменты, то из начала, середины или конца текста? Полные тексты, безусловно, предпочтительнее для лингвистики текста, но могут представлять проблему с точки зрения нарушения авторских прав. Кроме того, поскольку общий размер конечен, охват корпуса, содержащего полные тексты, не может быть сбалансирован, как, например, для корпуса, содержащего текстовые фрагменты одинакового размера. В результате «особенности индивидуального стиля или темы в силу конечного общего размера могут иногда проникать в обобщения» (Sinclair 1991). Астон и Бернард (Aston, Burnard 1998) считают, что «полнота» иногда может быть непримемлемой или проблематичной, в силу чего, если объектом исследования не является структурная организация текста или если отсутствует разрешение правообладателей, рекомендуется использвать фрагменты текстов. Как уже отмечалось, по-мнению Д. Байбера (Biber 1993), распространенные языковые явления распределяются достаточно стабильно и потому небольшие отрывки текста (напр., длиной в 2000 слов) бывают достаточными для

изучений таких явлений, в то время как распределение редких явлений – более изменчиво и для него потребуется более объемная выборка (Baroni 2009). При отборе текстов для корпуса необходимо следить за тем, чтобы выборки начала, середины и окончания текстов были сбалансированы.

Еще один вопрос, особенно относящийся к стратифицированной выборке, это соотношение и количество образцов для каждой категории текста. Для того, чтобы конечный корпус считался репрезентативным, количество образцов по текстовым категориям должно быть пропорционально их частотности и / или удельному весу в целевой совокупности. Тем не менее, отмечается, что, как и при определении целевой совокупности, такое соотношение бывает трудно определить объективно (Hunston 2002). Более того, критерии, используемые для классификации текстов по разным категориям, зачастую зависят от интуиции. Из-за этого репрезентативность корпуса должна рассматриваться как личное мнение, а не объективный факт. Например, для Брауновского корпуса комиссия экспертов определяла соотношения между 15 категориями текстов. Что касается количества текстов для каждой категории, Д. Байбер в своей работе *Representativeness in corpus design* (Biber 1993) доказывает, что десять образцов по 2000 слов каждый, как правило, являются достаточными.

5. **Однородность** (*homogeneity*) – корпус считается однородным, если содержащийся в нем материал был получен из одного источника или ограниченного числа источников. Например, однородность корпуса, состоящего из романов одного автора, будет очень высокой, корпуса статей из разных разделов одной и той же газеты за три года – менее высокой, а статей разных газет разных стран – еще меньшей. Наименее однородными будут корпусы, цель которых – отразить язык в целом, включая разнообразные типы текстов в устной и письменной форме (в частности, корпус *Bank of English* не является однородным вообще). В большнстве случаев язык однородных корпусов демонстрирует большую вариативность, чем разнородных. Если

перед исследователем стоит цель изучить конкретную узкую разновидность языка, тогда ему будет необходим однородный корпус, если же его цель – изучить, как данный язык функционирует в целом, то понадобится корпус с более широкой репрезентативностью.

6. **Машиночитаемая форма** (*machine-readable form*) – несмотря на то, что еще существуют немногочисленные корпусы в виде бумажных книг и аудиозаписей, большинство корпусов представлены в цифровом формате, что делает их обработку с помощью специальных компьютерных программ несравненно более легкой и быстрой.

8. **Образец для сравнения** (*standard reference*) – поскольку большинство крупных корпусов в наши дни доступны в сети Интернет и построены по приблизительно одним и тем же принципам, их можно использовать в качестве сравнения.

9. **Документирование** (*documentation*) – требование, предложенное Майком Нелсоном (Nelson 2000), состоящее в том, что каждый текст должен сопровождаться информацией о его жанре, типе текста, длине и пр.

Таким образом, перед тем, как приступить к созданию корпуса, исследователь должен ответить для себя на следующие вопросы:

1. Что должно являться основной единицей корпуса текстов (слово, морфема, предложение и т.п.)?

2. Каков должен быть объем корпуса текстов?

3. Какие источники должны быть представлены в корпусе текстов (устные или письменные, каких функциональных стилей и жанров, каких временных промежутков и т.п.)?

А.В. Зубов (2004) отмечает, что при решении последней задачи разработчики корпуса текстов обычно используют консультации специалистов по языкознанию и лингвостатистике либо метод анкет. Исходя из своего опыта исследований, специалисты определяют общий объем корпуса текстов, время издания текстов, число текстов и размер элементарной выборки, жанры отбираемых текстов и их

количество, число элементарных выборок каждого жанра. Так, при создании *The Brown Standard Corpus of American English*) группа консультантов-ученых определила его объем в 1 000 000 словоупотреблений. Было решено, что он должен состоять из 500 текстов по 2000 словоупотреблений каждый. Тексты должны быть взяты из произведений американских авторов, изданных в США в 1961 году. При этом было рекомендовано отобрать 15 письменных жанров (9 – информативная проза; 6 – художественная проза). Из каждого жанра было сделано от 6 до 80 элементарных выборок.

Метод анкет в сочетании с опытом специалистов был использован при создании корпуса текстов *The American Heritage Intermediate Corpus*. Специалисты, ориентируясь на заданное время создания корпуса, определили его объем в 5 000 000 словоупотреблений и рекомендовали включить в него лексику из 22 разделов (жанров) детской и юношеской литературы на английском языке. Для конкретизации текстов в 221 школу США были разосланы анкеты с просьбой указать, какие тексты желательно включить в корпус. После изучения анкет был составлен список из 19 000 названий книг. Из этого множества было отобрано 1045 текстов. На их основе было составлено 10 000 элементарных выборок по 500 словоупотреблений каждая.

Среди более частных проблем, решаемых перед созданием корпуса, ученые (в частности, В.В. Рыков) выделяют следующие:

1. Кто будет являться пользователем корпуса? (Индивид, группа, лингвистическое общество).

2. Какова логическая идея, положенная в основу корпуса?

3. Процедура отбора текстов в корпусе. Для разных целей по-разному:

а) обследование речевого материала;

б) сканирование текстов;

в) окончательное формирование, составление корпуса.

4. Аннотирование, индексирование словесного материала текста.

§ 4.4. Инструментарий работы с корпусом

Так же, как и сама корпусная лингвистика, программы обработки корпуса прошли несколько этапов развития, среди которых исследователи (McEnery, Hardie 2012; Anthony 2013) выделяют четыре поколения.

Первое поколение (1960-е – 1970-е гг.) работало на больших многопользовательских вычислительных системах (*mainframe*) и могли обрабатывать только наборы символов *ASCII* (латинский алфавит A-Z и ограниченный набор знаков, таких как знаки пунктуации и основные математические символы), в результате чего они были ограничены только англоязычными корпусами. Инструменты включали в себя отдельные функции – подсчет слов в тексте или конкордансы *KWIC*. Примерами программ первого поколения могут служить *Concordance Generator, Discon, Drexel Concordance Program, Concordance* и *CLOC* (последний использовался в знаменитом проекте *COBUILD* Бирмингемского университета под руководством Джона Синклера). Следует отметить, что идеи программ первого поколения лежат в основе современных программ.

Программы **второго поколения** (1980-е – 1990-е гг) также были ограничены набором знаков *ASCII* и имели узкую функциональность, однако они уже могли работать на персональных компьютерах, что позволяло ученым проводить собственные узкоспециализированные исследования, а также дало возможность преподавателям использовать корпусный анализ на занятиях языком – так называемое *обучение с помощью баз данных* (**data-driven learning**) (Johns 2002). Ко второму поколению относятся *Oxford Concordance Program, Longman Mini-Concordancer, Kaye Concordancer* и *MicroConcord*.

Большинство программ, используемых в наши дни, принадлежат к **третьему поколению**, первые версии которого появились еще в 1990-е годы, но многие продолжают разрабатываться и по сей день. Основным преимуществом этих программ является многофункциональность, включая основные

статистические методы, масштабируемость, позволяющая работать с более объемными корпусами, многоязычность, выходящая за рамки *ASCII*, а также удобный интерфейс для пользователей с небольшим опытом работы на компьютере. В число таких программ входят *WordSmith Tools, MonoConc Pro* и *AntConc*. Основным их недостатком является невозможность обработки очень больших корпусов (более 100 миллионов слов), что является особенно важным сегодня, когда есть возможность автоматической компиляции корпуса текстов Интернет-сайтов, объем которого может составлять несколько миллиардов слов. Другой проблемой является вопрос авторского права – сегодня издатели болезненно относятся к использованию своих материалов в исследовательских целях, что не позволяет обрабатывать их на персональных компьютерах.

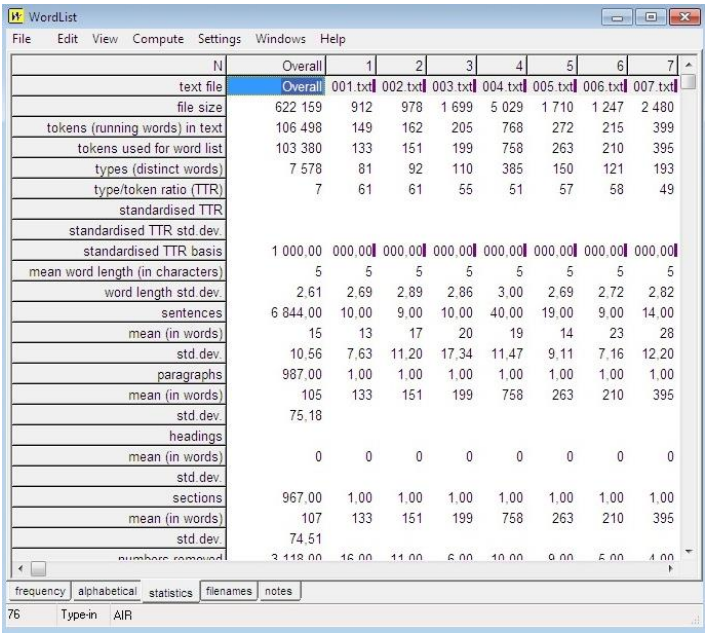
Преодолеть эти трудности призваны программы **четвертого поколения**, такие как *corpus.byu.edu, CQPweb, SketchEngine* и *Wmatrix*. Они обладают лучшей масштабируемостью за счет того, что корпус сохраняется в базе данных веб-сервера, а предварительная разметка данных ускоряет их поиск. Защита авторских прав обеспечивается за счет того, что пользователи не видят весь корпус, поскольку имеют доступ к нему только через специальный интерфейс, показывающий им в отдельно взятый момент времени лишь небольшой фрагмент при том, что есть возможность поиска по всему корпусу и получение результатов в виде частотных списков и конкордансов.

Несмотря на перечисленные достоинства, эти программы имеют ряд недостатков. Прежде всего это «чрезмерность»: когда, к примеру, пользователю нужно лишь провести простой анализ на небольшом корпусе, а для четвертого поколения еще до начала анализа требуется, чтобы данные были очищены, обработаны, переформатированы, размечены и загружены на сервер. Кроме того, чтобы получить доступ к серверу, пользователь должен зарегистрироваться, принять многочисленные лицензионные согласшения и оплатить подписку. Другая проблема состоит в том,

что эти программы не подходят для анализа корпуса конфиденциальных данных, таких как протоколы совещаний, вступительные экзамены, личные дневники, так как для анализа они должны быть загружены на внешний сервер. Третья проблема связана с тем, что программы четвертого поколения напрямую связаны с конкретным корпусом (обычно защищенным авторским правом) и не позволяют анализировать данные других корпусов. Четвертая проблема связана с третьей в том, что каждый новый (защищенный авторским правом корпус) выпускается вместе с новой программой, при отладке которой возникают многочисленные доступные только через Интернет интерфейсы, каждый со своим расположением управляющих кнопок и рабочими функциями. И наконец, программы четвертого поколения стирают границы между корпусными данными и инструментами для их наблюдения. В силу того, что эти инструменты хранят данные в размеченном виде на внешнем сервере, пользователи не имеют возможности наблюдать исходные данные собственными глазами, а только через пользовательский интерфейс веб-браузера, из-за чего исследователь часто забывает, что программа фильтрует данные, и полностью ей доверяет.

Как уже отмечалось, для работы с корпусом имеется ряд стандартных инструментов, доступных не только специалистам в сфере информационных технологий, но и обычным пользователям. Ниже рассмотрены основные из них – статистические данные, списки слов, частотные списки, ключевые слова, конкордансы, *n*-граммы и некоторые другие.

1. **Статистические данные** (*statistical data*) – большинство программ предлагают разнообразные статистические данные в отношении всего корпуса или отдельного текста. Ниже дан скриншот статистических данных программы *WordSmith Tools* (*Scott 1999*), где столбец 1 содержит данные о корпусе, а столбец 2 – данные о тексте:

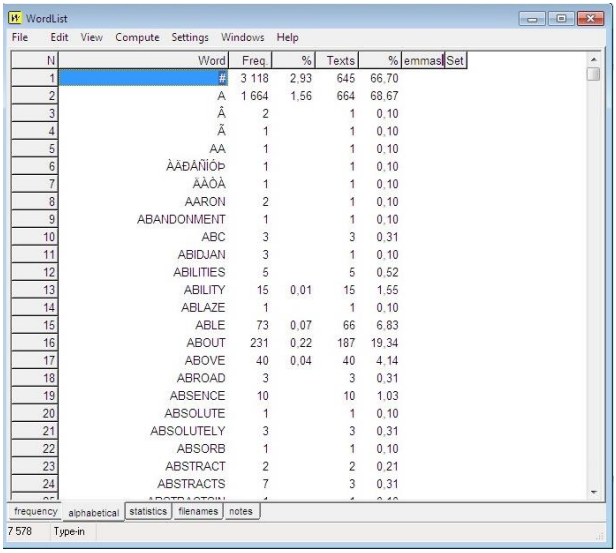


N	Overall	1	2	3	4	5	6	7
text file	Overall	001.txt	002.txt	003.txt	004.txt	005.txt	006.txt	007.txt
file size	622 159	912	978	1 699	5 029	1 710	1 247	2 480
tokens (running words) in text	106 498	149	162	205	768	272	215	399
tokens used for word list	103 380	133	151	199	758	263	210	395
types (distinct words)	7 578	81	92	110	385	150	121	193
type/token ratio (TTR)	7	61	61	55	51	57	58	49
standardised TTR								
standardised TTR std.dev.								
standardised TTR basis	1 000,00	000,00	000,00	000,00	000,00	000,00	000,00	000,00
mean word length (in characters)	5	5	5	5	5	5	5	5
word length std.dev.	2,61	2,69	2,89	2,86	3,00	2,69	2,72	2,82
sentences	6 844,00	10,00	9,00	10,00	40,00	19,00	9,00	14,00
mean (in words)	15	13	17	20	19	14	23	28
std.dev.	10,56	7,63	11,20	17,34	11,47	9,11	7,16	12,20
paragraphs	987,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
mean (in words)	105	133	151	199	758	263	210	395
std.dev.	75,18							
headings								
mean (in words)	0	0	0	0	0	0	0	0
std.dev.								
sections	967,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
mean (in words)	107	133	151	199	758	263	210	395
std.dev.	74,51							
numbers removed	3 118,00	16,00	11,00	6,00	10,00	9,00	5,00	4,00

Рис. 3. Пример статистической информации корпуса
(корпус деловой корреспонденции *VOBEC*, программа *WordSmith Tools*®)

2. **Списки слов** (*word lists*) – этот инструмент преобразует тексты в списки слов – последовательности знаков, разделенных пробелами. При этом различаются **индивидуальные слова** (*types*) и **словоупотребления** (*tokens*). В статистических данных может указываться соотношение индивидуальных слов к словоупотреблениям (*Type/Token Ratio, TTR*), отражающее лексическое разнообразие текста или корпуса. Списки могут быть представлены в алфавитном порядке или в порядке частотности. Алфавитные списки полезны тем, что позволяют увидеть формы одного слова или однокоренные слова (напр. *learn, learns. learned, learnt, learning, learner* и т.д.)[57]. Некоторые программы составляют даже обратные алфавитные списки, полезные для наблюдения за суффиксами и окончаниями.

[57] Более сложные программы, могут автоматически редуцировать списки слов в леммы, а некоторые даже различать их значения. Но такие программы как правило недоступны обычным пользователям.



N	Word	Freq.	%	Texts	%	emmas	Set
1	#	3 118	2,93	645	66,70		
2	A	1 664	1,56	664	68,67		
3	Â	2		1	0,10		
4	Ã	1		1	0,10		
5	AA	1		1	0,10		
6	ÀÄÐÁÑÏÓÞ	1		1	0,10		
7	ÄÀÒÀ	1		1	0,10		
8	AARON	2		1	0,10		
9	ABANDONMENT	1		1	0,10		
10	ABC	3		3	0,31		
11	ABIDJAN	3		1	0,10		
12	ABILITIES	5		5	0,52		
13	ABILITY	15	0,01	15	1,55		
14	ABLAZE	1		1	0,10		
15	ABLE	73	0,07	66	6,83		
16	ABOUT	231	0,22	187	19,34		
17	ABOVE	40	0,04	40	4,14		
18	ABROAD	3		3	0,31		
19	ABSENCE	10		10	1,03		
20	ABSOLUTE	1		1	0,10		
21	ABSOLUTELY	3		3	0,31		
22	ABSORB	1		1	0,10		
23	ABSTRACT	2		2	0,21		
24	ABSTRACTS	7		3	0,31		

Рис. 4. Пример алфавитного списка слов корпуса
(корпус деловой корреспонденции *VOBEC*, программа *WordSmith Tools*®)

3. **Частотные списки** (*frequency lists*) – показывают наиболее частоные единицы. Обычно именно с частотных списков исследователь начинает изучение корпуса. Частотная информация используется при составлении словарей, разработке учебных пособий и курсов. Тем не менее, как отмечает Сьюзан Ханстон (Hunston 2006), «информация о частотности не будет информативной, если она не будет сравнительной». **Ранговый порядок** или **рейтинг** в частотном списке (*rank order*) той или иной единицы сопоставляется в разных корпусах, языках, разновидности языка, жанрах и пр. Однако при проведении таких сопоставлений могут возникнуть трудности, вызванные разными объемами сопоставляемых корпусов. Чтобы решить эту проблему, различают **абсолютную частотность** (*raw frequency*) – фактическое количество употреблений отдельного слова в корпусе, и **нормализованную частотность** (*normalised frequency*) – количество употребление слова на миллион слов или на тысячу слов (Biber et al. 1998).

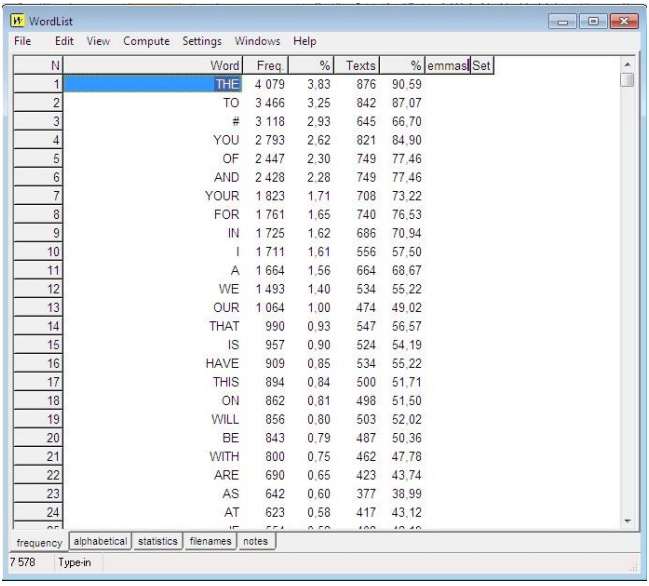


Рис. 5. Пример частотного списка слов корпуса
(корпус деловой корреспонденции *VOBEC*, программа *WordSmith Tools*®)

Существует интересная зависимость между рангом частотности слова и количеством слов в этом ранге, описанная законом Ципфа[58], который в упрощенном виде означает, что примерно половина слов встречается только один раз, четверть – два раза и так далее.

[58] Закон назван в честь **Джорджа Кингсли Ципфа** (*George Kingsley Zipf*), американского лингвиста (1902-1950), занимавшегося изучением статистических закономерностей в естественных языках, который примерно в 1935 г. впервые обратил внимание на это явление. Закон изучает частотность слов в естественном человеческом языке и то, как большинство наиболее частотных слов встречаются в два раза чаще, чем вторые по частотности, в три раза чаще, чем следующие и так далее до наименее частотного слова. Слово в позиции *n* встречается с частотностью *1/n* по отношению к самому частотному. Закон Ципфа имеет универсальный характер в естественном языке – он наблюдается как у детей в возрасте менее 32 месяцев, так и в специализированный лексике вузовских учебников. Исследования показали, что он применим почти к любому языку.

Так, в первом Брауновском корпусе список слов составлял 69002, из которых 35065 встречались только один раз. На другой стороне шкалы самое частотное слово (*the*) встречалось 69970 раз (одна десятая от общего количества), что было почти в два раза выше следующего по частотности слова (*of*) – 36410 раз (одна двадцатая). Иными словами, любой список слов (состоящий хотя бы из нескольких сотен слов) содержит небольшое количество высокочастотных единиц и длинный «хвост» редких единиц, при том, что приблизительно половину составляют *hapax legomena*, слова, встречающиеся только один раз (Scott, Tribble 2006).

К частотности тех или иных единиц в корпусе следует относиться с осторожностью – то, что они являются частотными в корпусе (особенно в специализированном), может не означать их частотности в языке в целом, а быть лишь связано с тематикой текстов корпуса.

4. **Ключевые слова** (*key words*) – это не обязательно самые частотные слова в корпусе, а те, которые демонстрируют статистически значимую частотность при сравнении с **контрольным** или **эталонным корпусом** (*reference / benchmark corpus*) большего или такого же объема. **Списки ключевых слов** (*key-word lists*) включают элементы, обладающие статистически значимой высокой частотностью – **положительные ключевые слова** (*positive key words*) или низкой частотностью – **отрицательные ключевые слова**

(*negative key words*). Ключевые слова обычно создаются автоматически (напр., инструмент *Keyword* программы *WordSmith Tools*) с помощью критерия логарифмического правдоподобия или критерия хи-квадрат. К ключевым словам обычно относятся имена собственные, служебные слова (которые могут указывать на стиль текстов) и знаменательные слова, связанные с тематикой текстов. Ниже показаны ключевые слова из устной части корпуса *BNC-OU* (4-миллионного подкорпуса из *British National Corpus*).

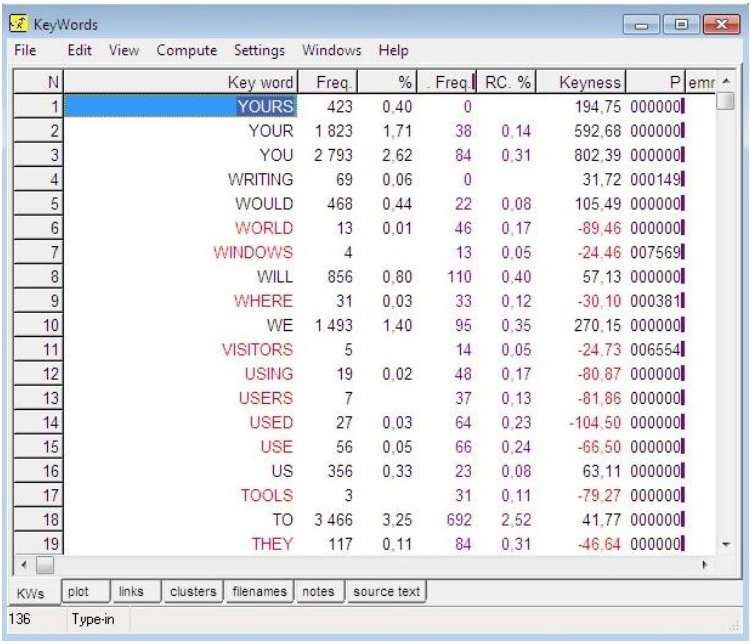


Рис. 6. Пример списка ключевых слов корпуса деловой корреспонденции *VOBEC* по сравнению с подкорпусом *BNC-OU* корпуса *Britich National Corpus* (программа *WordSmith Tools*®)

5. **Конкордансер** (*concordancer*) – компьютерная программа, в автоматическом режиме строящая **конкордансы** (*concordances*) – искомые слова в окружающем контексте (*Key Words In Context*, сокращенно – *KWIC*). Контекст не представляет собой законченное

предложение или абзац, а задается количеством слов справа и слева, которые могут регулироваться пользователем. При необходимости можно затребовать более широкий контекст. Поиск слова/слов производится с помощью **запроса** (*query*), в котором пользователь задает нужные ему параметры и единицы. Ниже приведен фрагмент списка конкордансов, составленный с помощью *WordsmithTools*.

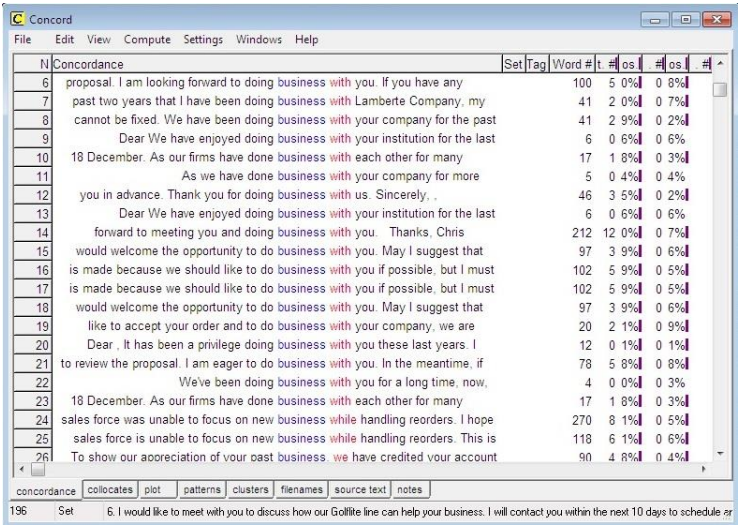


Рис. 7. Пример списка конкордансов слова *business*
(корпус деловой корреспонденции *VOBEC*, программа *WordSmith Tools*®)

Полученные данные можно сортировать в зависимости от задачи. Чаще всего конкордансы располагаются в том порядке, в котором они встречаются в тексте, однако программа позволяет сортировать левый или правый контекст ключевого слова (до пяти слов) в алфавитном порядке. Кроме того, в качестве элемента запроса может выступать не одно слово, а цепочка слов, а с помощью специальных символов можно осуществлять поиск форм одного слова, например, запрос *sort** предлагает формы *sort, sorts, sorted, sorting, sortable*, а запрос *r*t* покажет все слова корпуса, начинающиеся с ‘*r*’ и оканчивающиеся на ‘*t*’ – от *rat* до *restaurant*.

Данный инструмент позволяет исследователю увидеть сочетания, образующиеся с тем или ином словом, что, в свою очередь, демонстрирует более тесную связь между грамматическими и лексическими моделями, чем было принято считать раньше. Связи между грамматической структурой и лексическими единицами могут анализироваться с точки зрения семантических характеристик этих единиц, так называемой «семантической просодии» (*semantic prosody*), когда определенные структуры тяготеют к определенным типам значания, например, положительные или отрицательные обстоятельства (Sinclair 1991). В частности, в работе Анны О'Кифф и Майкла МакКарти (O'Keeffe et al. 2007) приводится объемное корпусное исследование страдательных конструкций со вспомогательным глаголом *get* (напр., *he got arrested*), которое показывает, что эта конструкция обучно употребляется для описания неблагоприятных ситуаций, что выражается через семантику таких глаголов, как *killed, sued, beaten, arrested, burgled, intimidated, criticised* и др. Изолированно эти глаголы встречаются гораздо реже, чем в этой конструкции.

Конкордансеры являются очень эффективным исследовательским инструментом, поскольку позволяют объединить большое количество употреблений одного элемента и увидеть их в исходном контексте, что чрезвычайно полезно как при проверке гипотезы, так и при выдвижении новой. Однако, как отмечает Сьюзан Ханстон (Hunston 2002), при проверке гипотезы нужно быть внимательными по отношению к фактам, которые ее не подтверждают, – при необходимости лучше пересмотреть исходную гипотезу, чем отбросить противоречащие ей факты.

6. **N-граммы** (*n-grams / lexical bundles / chunks / clusters*) – непрерывные последовательности элементов (количество элементов обозначается буквой *N*), встречающиеся в тексте или корпусе. Такими элементами могут быть фонемы, слоги, буквы или слова. Чаще всего анализируются лексические N-граммы, составляющие словосочетания или клише (напр., *you know, in the, there was a, one of the*). Обычно они группируются на основе количества слов, при этом

их частотность резко падает в зависимости от количества этих слов, например, в корпусе, состоящем из 5 млн слов (Carter, McCarthy 2006) и насчитывающем 45015 двухсловных сочетаний, было обнаружено только 31 шестисловное сочетание, встречающееся 20 и более раз. Это показывает, что, несмотря на клишированность языка, он остается уникальным и своеобразным.

Инструменты построения *N*-грамм есть в программе *WordsmithTools* (см. рисунок ниже), *AntConc*, бесплатной программе *kfNgram*, функционирующая в среде *Windows* (http://miniappolis.com/KWiCFinder/kfNgramHelp.html), а также с помощью поискового онлайн-сервиса компании *Google*, *Google Ngram Viewer* (https://books.google.com/ngrams), охватывающего более 5,2 млн книг, опубликованных с XVI в. и собранных в сервисе *Google Books*.

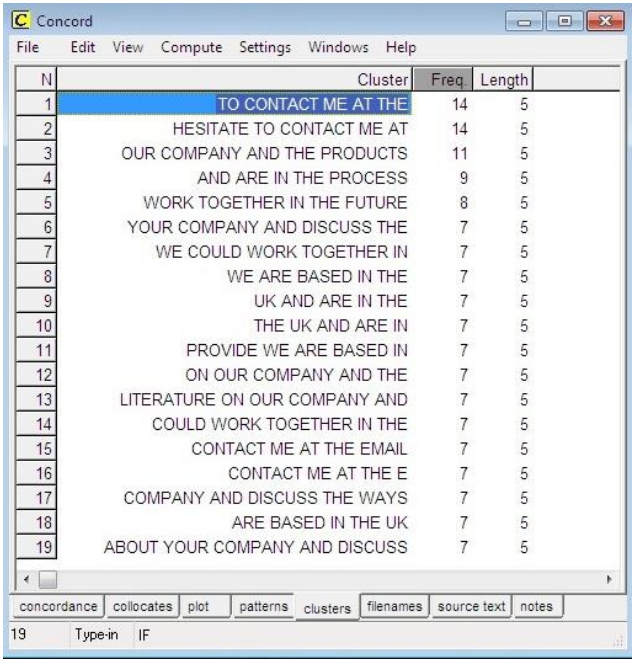


Рис. 8. Пример построения *N*-грамм (для 5 слов) в корпусе (корпус деловой корреспонденции *VOBEC*, программа *WordSmith Tools*®)

7. **График дисперсии** (*dispersion plot*) – показывает распределение заданных лексических или грамматических единиц в тексте – равномерно по всему тексту или преимущественно в одной части (начале, середине или конце), например, фраза *Once upon a time* будет находится в левой части графика, а *happily ever after* – в правой. Распределение той или иной единицы в тексте или ряде текстов может дать интересную информацию о структуре текста и функционировании данных единиц. Ниже приведен график дисперсии слова *sincerely* (программа *WordsmithTools*) в корпусе деловых писем VOBEC, где, как можно было предполагать, оно расположено в заключительной части текста.

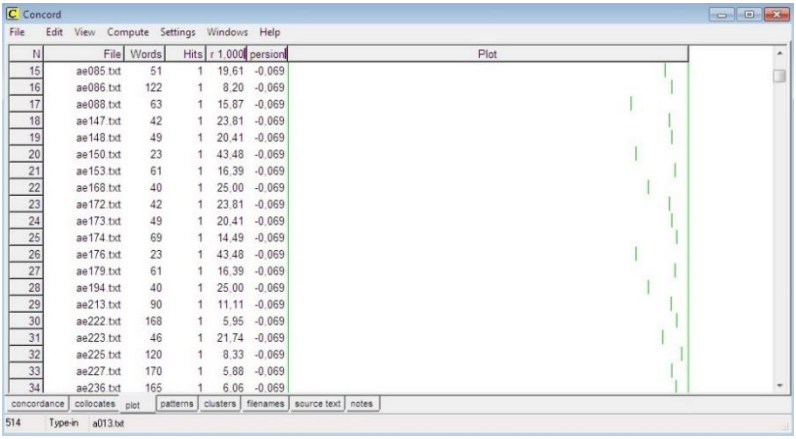


Рис. 9. Пример графика дисперсии слова *sincerely* в корпусе деловых писем (корпус деловой корреспонденции *VOBEC*, программа *WordSmith Tools*®)

Таковы основные базовые инструменты, имеющиеся в распоряжении исследователя-лингвиста для компьютерного анализа корпуса. Более сложные и продвинутые инструменты могут потребовать знаний в области математики, основ программирования и статистики[59].

[59] Основные принципы и роль статистики в научных исследованиях в целом и в корпусной лингвистике в частности изложены в книге Вацлава Брезиня *Statistics in Corpus Linguistics: A Practical Guide* (Brezina 2018).

5. Практика корпусных исследований

§ 5.1. Корпусные грамматические исследования

Обращение к речи является краеугольным камнем корпусной лингвистики. Тексты и высказывания в составе корпуса относятся к разным функциональным стилям – художественная литература, журнальные и газетные статьи, радио– и телепередачи, разговорная речь[60]. Изучение корпусных материалов дает исследователям объективную информацию о разных свойствах современного английского языка, в том числе, грамматических. Первым опытом подобного рода было исследование, названное *Survey of English Usage* и осуществлявшееся под руководством заведующего кафедрой английского языка Лондонского университета профессора Р. Кверка[61]; оно имело своей целью дать полное и непредвзятое описание английского языка. Материалы, выводы и наблюдения стали основой для «Грамматики современного английского языка», вышедшей в свет в 1972 году и явившейся событием в истории английской филологии, значительно отличаясь от предшествующих ей трудов в области описания грамматического строя английского языка.[62]

В 1999 году вышла новая книга по грамматике, основанная на корпусных исследованиях, – *Longman Grammar of Spoken and Written English (LGSWE)* под ред. Д. Байбера, С. Джохансона, Дж. Лича, С. Конрад и Э. Финегана.[63] Она во многом повторяет терминологию и общий подход к грамматике своей предшественницы, однако ее

[60] “ … considerable thought is given to the selection of material, so that, in the most general case, the corpus can stand as a reasonably representative sample of the language as a whole.” Crystal 1995:438.

[61] Подробнее об этом и других корпусах см.: Crystal 1995:438-441; Sinclair 1991; Biber 2004.

[62] Quirk, Greenbaum et al. 1972.

[63] Biber et al. 1999. Ср. с другими корпусными исследованиями грамматики: Quirk, Greenbaum, Leech, Svartvik 1985; Greenbaum, Quirk 1990; Sinclair 1990; Leech, Svartvik 1994; Greenbaum 1996; Carter, McCarthy 2006.

иллюстративный материал и грамматический анализ различных разновидностей английского языка несопоставимо шире. Материалом для данной книги послужил корпус *Longman Spoken and Written English Corpus* объемом около 40 026 000 слов, содержащий 37 244 текста.

Сами составители новой грамматики видят ее отличие от традиционных грамматик прежде всего в том, что последние были заняты в основном структурным анализом, систематизируя и описывая форму и значение грамматических конструкций, не уделяя должного внимания их реальному функционированию в устной и письменной речи. Их описательный характер был вызван тем, что само понятие грамматики тогда понималось как изучение структур. Среди причин недостаточного учета функциональных особенностей языковых явлений отмечается отсутствие информации о том, что действительно имеет место в реальных контекстах речеупотребления.

Для изучения грамматических явлений, их свойств и закономерностей функционирования авторы *LGSWE* выбрали четыре основных разновидности английского языка: разговорную речь или бытовой диалог (*Conversation*); художественную литературу *(Fiction*), – представленную как британским, так и американским вариантами английского языка; газетные тексты (*Newspaper language*) и академическую прозу (*Academic prose*). Два вспомогательных регистра представлены небытовой устной речью (*Non-conversational speech*), такой как лекции, учебные занятия, проповеди, аукционы, медицинские консультации, интервью, собрания, радио- и телевизионные передачи и т.п., а также нехудожественной прозой (*Non-fiction prose*), включающей в себя популярную литературу по разным темам – экономика, образование, искусство, кулинария, история, увлечения, литературоведение, лингвистика, философия и т.д. Здесь же даются тексты по теме «бизнес» (*Business Subject*), представляющие такие области, как финансы, маркетинг, управление персоналом, коммерция и пр.

Каждая из обследуемых разновидностей речи обозначена термином *register*, а каждый используемый образец называется *text*.

Тексты существуют в письменной и устной форме; они воспринимаются дискурсивно с учетом лингвистических, ситуационных, социальных, психологических и прагматических факторов, влияющих на их интерпретацию. Собрание устных и письменных текстов, организованных по регистрам и закодированных с целью последующего дискурсивного анализа, и составляет «корпус».

По мнению авторов *LGSWE*, именно данные четыре регистра – *Conversation, Fiction, Newspaper language, Academic prose* – демонстрируют наибольшую широту охвата ситуативного и языкового варьирования английского языка. На одном полюсе континуума находится бытовой диалог, которым владеют практически все носители данного языка. На противоположном полюсе – академическая проза, являющаяся наиболее специализированной: лишь ограниченная часть носителей языка регулярно читает научные тексты, еще меньшая часть способна их производить. Между этими двумя полюсами расположены газетные тексты и художественная литература. Они представляют собой письменную разновидность речи, однако являются достаточно распространенными и не столь специализированными, как академическая проза. Большинство людей хотя бы время от времени читают произведения художественной литературы и газетные статьи. Основное различие между этими двумя регистрами видится составителями корпусной грамматики следующим образом: художественная литература ориентирована на эстетическое воздействие и развлечение читателей; газеты призваны информировать. Последние, как правило, претендуют на объективность и непредвзятость изложения, а потому их тексты построены таким образом, чтобы личное мнение автора не было бы выражено слишком явно.

Бытовые диалоги кардинальным образом отличаются от газетных новостей. Это, прежде всего, связано с тем, что они представлены в устной форме и производятся в режиме «онлайн». Выбор слов и грамматических конструкций осуществляется по ходу развертывания

диалога – отсутствует какое-либо планирование или редактирование. Другой особенностью бытового диалога является его персонифицированный и интерактивный характер. Он представляет собой результат деятельности обоих участников и обычно связан с их собственной жизнью, проблемами, интересами. Кроме того, диалог происходит в рамках одного и того же пространственного и временно́го контекста и предполагает наличие у коммуникантов общих фоновых знаний.

Художественная литература занимает промежуточное положение между бытовым диалогом и остальными двумя регистрами, поскольку содержит диалоги персонажей, представляющие, хотя и искусственные, но образцы устной разговорной речи.

Помимо ситуационных отличий представленные в *LGSWE* четыре регистра различаются по своим диалектным характеристикам. Бытовой диалог и газетные тексты носят «региональный», «локальный» характер в том смысле, что отражают особенности диалекта того региона, где они были созданы: собеседники преимущественно используют в разговоре один и тот же региональный или социальный диалект. Аналогичным образом, газеты (особенно местные) обычно ориентированы на один конкретный регион или государство и таким образом отражают на письме национальные отличия между, например, британским и американским вариантами английского языка. Научные и художественные тексты, как подчеркивается составителями *LGSWE*, можно охарактеризовать как «глобальные» в том смысле, что они ориентированы на широкую международную аудиторию и потому меньше зависят от национальных диалектов своих авторов.

Исследование корпуса *Longman Spoken and Written English Corpus* показывает, что вариативность явлений часто носит весьма системный характер: пользователи языка делают выбор в области морфологии, словаря и грамматики в зависимости от ряда лингвистических и экстралингвистических факторов. Важными компонентами ситуационного контекста являются цель коммуникации, физическая разновидность (устная или письменная), обстоятельства

речепроизводства, а также разнообразные демографические характеристики говорящего / пишущего. Образцы вариативности, вызванные данными факторами, могут анализироваться по отношению к двум основным языковым разновидностям, выделяемым в составе *LGSWE,* – «диалектам» (*dialects*), связанным с различными группами говорящих, и «регистрам» (*registers*), определяемым ситуационно.

Каждому регистру присущ свой набор грамматических явлений. Так, например, регистр бытового диалога[64] отличается следующими грамматическими особенностями:

Некоторые монолексемные глаголы (такие, как *think, know*) и фразовые глаголы (*come on* и *get in*), а также ряд модальных глаголов (e.g. *can, will, would*) гораздо шире представлены в текстах бытовых диалогов, чем, к примеру, в текстах новостей.

Частотное употребление сложноподчиненных предложений с придаточными изъяснительными, вводимыми союзом *that*, причем союз чаще всего опускается (*I think Stuart's gone a bit mad!*). Из других типов придаточных предложений широко представлены придаточные условия, причины и времени.

Зачастую само понятие «предложение» оказывается весьма условным. Если руководствоваться пунктуацией, маркирующей конец предложения (точка, вопросительный знак, восклицательный знак), то большинство предложений имеют небольшую длину. Сюда относятся краткие ответы (*Yeah.* | *Okay.*), незаконченные высказывания, вызванные изменением намерения говорящего (e.g. *No, two. I'll… I'll… How much are those? Okay, so we need to put – I'm confused now*). Ситуационно обусловленный эллипсис связан с опущением единиц, имеющих низкую информативную нагрузку, чаще всего подлежащего (*Got a pen? Didn't know it was yours?*).

---

[64] Регистру бытового диалога, в отличие от других, в книге посвящена отдельная глава, что авторы объясняют следующим образом: “This is not only because it represents the spoken language in contrast to the other three registers, but also because the grammar of conversation has been little researched until recently, when the advent of sizeable computer corpora have made such research feasible for the first time.” (Biber et al. 1999:1038). В связи с грамматикой устной речи см. Brazil 1995.

Многочисленные сокращенные формы (*I'll.* | *Here's.* | *I'm.* | *That's*.).

В силу своей интерактивной природы бытовые диалоги отличаются широким использованием местоимений 2-го лица, т.е. адресата (*you*), а также прямых вопросов (e.g. *Do you want that one?* | *How much are those?*) и императива (e.g. *Put that towards her bill.* | *Just ring it all up together*.).

Говоря о своих собственных мыслях и чувствах, о событиях, произошедших с ними в настоящем или прошлом, участники бытового диалога часто используют местоимения 1-го лица (*I, we*) – например, *Hey, we like to see those.* | *I want this too.* | *Yes, I do.* | *Okay, so we need to put – I'm confused now.*

В бытовом диалоге преобладают глагольные формы настоящего времени, а также модальные словосочетания, отражающие сиюминутный характер общения собеседников (*Here's a twenty.* | *They can all go.*) или их эмоциональное состояние в данный момент (*We like…* | *I want…*).

С точки зрения соотношения семантических и функциональных лексических единиц, тексты бытовых диалогов характеризуются большой представленностью функциональных слов и меньшей – семантических слов (всего 41% по сравнению с 63% в текстах газетных новостей), что объясняется меньшей выраженностью информативного аспекта в данной функциональной разновидности языка. Преобладающими частями речи являются глагол и наречие, а также местоимения, замещающие существительные, поскольку референта местоимения собеседникам легко установить с опорой на ситуацию общения. По сравнению с другими регистрами в бытовом диалоге меньше всего употребляются артикли, что связано с низкой частотностью определяемых ими существительных.

Грамматические особенности регистра газетных новостей как письменной разновидности языка зачастую противоположны бытовому диалогу:

Для газетных текстов, отличающихся наличием информативного фокуса, характерно употребление существительных, обозначающих

людей или явления в мире. Ввиду требования компактности изложения информации новостной дискурс демонстрирует тенденцию к «плотности», т.е. насыщенности каждого текста существительными, прилагательными и предложными сочетаниями (e.g. *He said that as long as there is a demand for drugs, dealers and producers will find ways to meet that demand*.). При этом все формы даются полностью, без сокращений.

Придаточные изъяснительные, вводимые союзом *that*, чаще всего используются для передачи речи третьих лиц, при этом союз сохраняется (e.g. *They denied that they had adopted a plan for guerrilla warfare.*).

Все предложения отличаются законченностью и полнотой; многие имеют большую длину и сложную структуру (e.g. *Wright said that he would serve as Speaker until the Democratic Caucus chooses his successor at a session Tuesday*.). Наиболее частотным типами придаточных предложений являются придаточные времени, цели и причины.

Газетные тексты обычно обращены не к конкретному адресату, а к широкой читательской аудитории, поэтому в них отсутствуют местоимение 2-го лица (*you*), прямые вопросы и императивные предложения.

Тон газетных новостей должен быть максимально объективным, обезличенным, дистанцированным от личности автора, что ведет к отсутствию в них местоимений 1-го лица (*I, we*). В предложениях, передающих чьи-либо слова, роль подлежащего обычно берут на себя местоимения 3-го лица (*he, she, they*) или имена собственные (*Charles Hernu, Wright, McCartor*).

С точки зрения словарного состава, в текстах новостей существенно возрастает, по сравнению со служебными словами, процент полнозначных лексических единиц (63%), что объясняется высокой информативной нагруженностью таких текстов.

В силу основной задачи новостных текстов – сообщить о том, что уже произошло, почти все глаголы употребляются в формах

прошедшего времени (e.g. *He declared that she had researched projects, studied stocks and visited drilling sites to guide the firm's investments.*).

Различия между бытовыми диалогами и газетными новостями в значительной мере объясняются тем, что они принадлежат к разным – устной и письменной – разновидностям речи.

С учетом вышесказанного целесообразно продолжить рассмотрение двух других регистров в составе *LGSWE* и обратиться к академической прозе, особенности которой могут быть обобщены следующим образом:

В этом регистре меньше всего представлены глаголы, поскольку основное внимание сосредоточено не на событиях, а на отношениях между абстрактными понятиями, в связи с чем научные тексты содержат сложные именные и предложные словосочетания. Этим же объясняется высокая частотность глагола-связки *be*, который служит для соединения слов и словосочетаний, а не для указания на действие. Значительно большую частотность имеют существительные и связанные с ними артикли (в рассматриваемом регистре выявлен самый высокий процент употребления артиклей). По сравнению с другими регистрами научные тексты содержат больше всего прилагательных, которые выступают в функции определения существительных с одной стороны и в составе именного сказуемого с другой.

Академическая проза демонстрирует самую низкую частотность употребления личных местоимений. Что касается местоимения *we*, то оно может относиться к одному автору, группе авторов, автору и читателю или к людям вообще (e.g. *When we start talking we often cease to listen*.). Из указательных местоимений наиболее употребительными являются *this/these*, делающие отсылку к непосредственно предшествующему контексту.

Формы страдательного залога составляют 25% от общего числа личных форм глагола, что является наибольшим процентом по сравнению с другими регистрами. Это объясняется необходимостью сместить внимание с исполнителя действия на само действие (e.g. *They were based on his book The Principles of Quantum Mechanics*.).

Академическая проза отличается широким использованием модальных глаголов *may, must* и *should*. Основным значением является «логическая возможность» (e.g. *The only problem may be that the compound is difficult to remove after use*.).

Употребление сравнительных форм прилагательных объясняется ролью сопоставления как средства осознания и экспликации действительности: через сопоставление или противопоставление каких-либо явлений легче объяснить их природу (e.g. *Internodes are longer and sheaths relatively and progressively shorter than the internodal length*.).

Еще одной особенностью академической прозы является широкое использование придаточных обстоятельственных цели, условия и уступки; последние встречаются в научной речи чаще, чем в остальных регистрах (e.g. *It is possible to separate one from the others, though in certain situations one aspect may be more involved*.).

Регистр художественной литературы, четвертый из тех, что стали объектом детального рассмотрения в *LGSWE*, содержит диалоги персонажей, что сближает его с регистром бытового диалога. Есть у этого регистра и такие характеристики, которые отличают его от трех других ключевых функциональных разновидностей в составе *LGSWE,* например:

Употребление форм прошедшего времени. В действительности большинство художественных повествовательных текстов написано с использованием исключительно форм прошедшего времени, настоящее время используется преимущественно в речи персонажей. Частотность форм прошедшего времени совершенного и продолженного видов в регистре художественной литературы намного выше, чем в других письменных регистрах, таких как газетные новости, где также используются формы прошедшего времени (e.g. *When he returned the priest had already used the special needle-sharp quill and ink*.).

Использование сравнительных конструкций с союзом *as* как средства образного воссоздания реальности. В составе метафоры или гиперболы они апеллируют не столько к разуму читателя, сколько к

его воображению (e.g. *The sea came in with black waves as high as church towers and mountains*.).

Присутствие в предложении различных видов обстоятельств. С помощью обстоятельств в художественной литературе создается вымышленный мир, они служат для описания обстановки, характеров, действий и т.п. (e.g. *Color flamed vividly in a profusion of variegated reds and oranges intermingled with magenta and purple*.).

Использование придаточных времени, образа действия и сопутствующего действия (*supplementive clause*) для описания событий, действий и т.п. (e.g. *He began to puff his pipe, no doubt arranging his opinion in his mind*.).

LGSWE дает детальное представление об онтологии того или иного грамматического явления (*target feature*) во всех четырех регистрах – *Conversation, Fiction, Newspaper language, Academic prose*. Для иллюстрации сказанного целесообразно обратиться к страдательному залогу.

Наибольшая частотность страдательного залога наблюдается в регистре академической прозы (18 500 раз на 1 млн. слов); относительно меньше этих форм в газетных текстах (12 000 раз на 1 млн. слов). В основном страдательный залог представлен в регистрах с наименьшим количеством глаголов в личных формах (25% от всех глаголов в личной форме в академической прозе и 15% – в газетных текстах). В бытовом диалоге пассивные конструкции составляют только 2% всех глаголов в личной форме.

Одной из основных функций пассива в академической прозе является избежание упоминания агенса (лица, выполняющего действие, выраженное глаголом) и придание теме статуса пациенса. Иногда в страдательном залоге бывают написаны целые абзацы:

Three communities on a blackish marsh of the Rhode River, a sub-estuary of the Chesapeake Bay, were exposed to elevated carbon dioxide concentrations for two growing seasons beginning in April 1987. The study site and experimental design are described in Curtis et al. (1989a). One community was dominated by the perennial carbon 4 grass spartina patens … (ACAD)

Зачастую страдательный залог в академической прозе позволяет не упоминать конкретного исследователя (когда речь идет о каких-либо аспектах научных методов или анализа), а иногда придает некую объективность, отстраненность от описываемых фактов.

В газетных новостях употребление страдательного залога вызвано несколько другими причинами. Часто основной темой заметки является событие, повлиявшее на какое-либо лицо или учреждение. Нередко исполнитель действия устанавливается с опорой на конкретную ситуацию. Кроме того, он может быть неинтересен, или ранее уже упомянут. Вследствие этого, а также из-за стремления журналистов сэкономить место для максимальной подачи новой информации, исполнитель действия, как правило, не называется:

Doherty was arrested in New York in June. (NEWS)

В бытовом диалоге, напротив, основное внимание обращено на человека, его действия, мысли, эмоционально-психологическое состояние. Поэтому исполнитель действия (зачастую это сам говорящий) чаще всего называется. Однако несколько глаголов в страдательном залоге являются более употребительными в бытовом диалоге, чем в письменных регистрах. Самый частотный из них входит в состав устойчивого выражения can't be bothered:

I can't be bothered really. (CONV)

Значительное влияние на выбор между действительной и страдательной конструкцией оказывают лексические факторы – одни глаголы чаще принимают формы страдательного залога, другие – действительного. Некоторые, такие как *to be born* (в значении «родиться») и *to be reputed*, встречаются исключительно в форме страдательного залога:

Brandon Lee was born in Oakland, California. (NEWS)

The deal is reputed to be worth £1m. (NEWS)

Другие глаголы (*to be based on, to be deemed, to be positioned и to be subjected to*) могут употребляться как в действительном, так и в страдательном залоге, однако последний составляет более 90% употреблений.

Ряд глаголов в страдательном залоге чаще употребляется в регистре академической прозы (*to be analysed, to be calculated, to be collected, to be measured, to be tested, etc*.), другие – в газетных новостях. При этом зачастую речь идет о каких-то неприятных событиях, произошедших с людьми (*to be accused, to be jailed, to be killed, to be wounded, to be shot, to be injured, etc*.)

Говоря о проблеме выбора, можно проанализировать функционирование двух синонимичных конструкций с точки зрения не только их общей употребительности, но и частотности в каждом конкретном регистре. Возьмем, к примеру, форму притяжательного падежа существительных и синонимичное словосочетание с предлогом *of* [65]. В целом данные корпуса показывают, что во всех регистрах преобладают словосочетания с предлогом *of*. При этом каждый регистр имеет свои особенности. Так, в бытовом диалоге указанные конструкции представлены менее всего. Наибольшую частотность притяжательного падежа демонстрируют газетные тексты, а предложных словосочетаний – академическая проза.

Распределение данных грамматических явлений по регистрам связано с общей частотностью существительных. Тот факт, что в бытовом диалоге существительные встречаются гораздо реже, снижает и частотность рассматриваемых словосочетаний. В академической прозе и газетных новостях существительные представлены значительно шире, что объясняет и более высокую частотность словосочетаний с притяжательным падежом и предлогом *of*. С этой точки зрения регистр художественной литературы занимает промежуточное положение.

Высокая частотность притяжательного падежа в газетных новостях, возможно, объясняется тем, что он является одним из

[65] Из анализа исключаются те словосочетания, которые не являются альтернативой притяжательному падежу – устойчивые выражения с глаголами и прилагательными (*accused of, afraid of, etc*.), сложные предлоги (*in front of, because of, etc*.), словосочетания с числительными, местоимениями, собирательными существительными, а также существительными со значением единицы измерения, количества, разновидности и пр. *(one of, some of, a piece of, a herd of, a box of, types of, etc*.).

способов сжатия информации, что очень важно для лаконичности новостных заметок.

В художественной литературе притяжательный падеж употребляется значительно чаще, чем в академической прозе, что объясняется характером существительных, представленных в этом регистре (в основном существительные, обозначающие людей).

Преобладание словосочетаний с предлогом *of* над словосочетаниями с притяжательным падежом, вероятно, обусловлено предпочтительным использованием менее компактных структур с целью достижения большей прозрачности высказывания. Кроме того, данный факт отражает современное состояние исторического сдвига в сторону предложного сочетания, который постоянно продолжается, начиная с древнеанглийского периода, когда преобладала флективная форма притяжательного падежа.

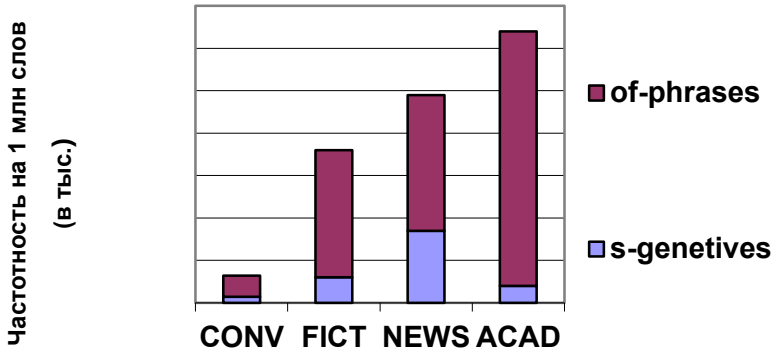


Рис. 10. Распределение сочетаний с предлогом *of*
и притяжательным падежом по четырем регистрам

Порядок расположения регистров при их сопоставлении (бытовой диалог → художественная литература → газетные тексты → академическая проза) также носит неслучайный характер. Авторы *LGSWE* отмечают, что во многих (хотя и не во всех) случаях возрастание (или, наоборот, убывание) частотности употребления тех

или иных грамматических явлений происходит именно в этой последовательности.

Используемый в *LGSWE* корпусный подход позволяет продемонстрировать реальное функционирование английской грамматики в коммуникативной деятельности говорящих и пишущих. Каждый регистр обладает своими собственными грамматическими особенностями, а каждое грамматическое явление по-разному проявляет себя в различных регистрах.

§ 5.2. Корпусные лексические исследования

В плане изучения слова и его семантики корпусная лингвистика эффективно помогает решать следующие вопросы (Biber 2004):

1. Каковы основные значения слова? Корпусная лингвистика позволяет определять значения слов, анализируя их естественный контекст, а не полагаясь на интуицию или выборочные собрания цитат.

2. Какова частотность слова по сравнению с другими словами? Такого рода исследования позволяют выявить наиболее частотные или редкие слова.

3. Какие у данного слова экстралингвистические связи (например, относящиеся к регистру речи, историческому периоду или диалекту)? Решение подобных вопросов позволяет определить модели употребления лексических единиц в различных разновидностях языка, например, отличия, связанные с их использованием в разных регистрах или с изменением во времени.

4. С какими словами чаще всего встречается данное слово и какова дистрибуция данных коллокаций по разным регистрам? Такие исследования обращают внимание на модели сочетаемости слов друг с другом – коллокации.

5. Какова дистрибуция различных значений и употребления слова? Здесь решаются такие вопросы, как:

- Насколько частотны различные значения и грамматические функции слова?
- Имеется ли системность в распределении этих значений и функций по регистрам?

Необходимо также отметить, что корпусный подход помогает противопоставить интуицию или предписания носителей языка эмпирическим данным по реальному функционированию языка. Такое противопоставление не просто представляет интерес в плане выявления ошибочных представлений о языковых явлениях, но позволяет создавать словари, учебные пособия и методические

материалы для студентов всех уровней, основанные на более точных данных. Корпусные инструменты обнаруживают реальное функционирование языка в естественных контекстах и выявляют те словоупотребления, с которыми студенты наиболее вероятно смогут столкнуться в своей работе, вместо того, чтобы полагаться на мнения авторов книг по поводу значения и употребления слов.

Следует иметь в виду, что лексический анализ требует использования весьма объемных корпусов, поскольку значения слов и коллокационные модели обладают гораздо меньшей частотностью, чем грамматические модели.

5.2.1. Анализ значения слова

Как показали корпусные исследования, реализация любого языкового элемента зависит от окружающего его контекста. Детали языкового выбора говорящего в определенном отрезке дискурса в значительной степени обусловлены соответствующими языковыми решениями. Следовательно, любой пример, лишенный своего текстового окружения или придуманный в целях иллюстрации того или иного значения слова, не может считаться полноценным. Как справедливо отмечает Дж. Синклер, «в качестве примера языкового использования никогда нельзя приводить какую-либо комбинацию слов, употребление которой не может быть подтверждено реальными речевыми фактами» (Sinclair 1991).

Одним из преимуществ корпусных исследований является то, что корпус помогает продемонстрировать все контексты данного слова. Из этих контекстов далее становится возможным определить различные значения слова. Раньше для этих целей использовались цитаты. Основное отличие состоит в том, что цитаты представляют собой только те контексты, на которые обратил внимание читательский глаз (иногда представляющие неординарные случаи). Корпусный анализ, напротив, предоставляет исчерпывающий перечень всех употреблений и контекстов.

В наши дни существует множество программ, помогающих производить лексический анализ слов, в частности, списки конкордансов, где каждое употребление слова представлено отдельной строкой, в середине которой расположено выбранное слова, а слева и справа – его контекст. Такой экран называется *KWIC – Key Word in Context*. Иногда бывает необходимо увидеть более протяженный контекст, и тогда конкорданс *KWIC* может показывать несколько строк. Ниже представлен список конкордансов слова *deal* в корпусе *LOB.*

and secret plans to	**deal**	with the mass sit-down	1
of companies and put one	**deal**	through each. Mr.	2
property			
. In particular, a good	**deal**	of concern has been	3
hangs a tale – and a great	**deal**	of money. Neville	4
where his new measures to	**deal**	with Britain's	5
just a matter of working a	**deal**	harder before we	6
good		really	
. "I'm mixed up in a	**deal**	involving millions	7

Даже из этого короткого перечня можно увидеть, что *KWIC* демонстрирует различные значения слова: в первой и пятой строках это «иметь дело с проблемой», такой как *mass sit-down*; во второй и седьмой строках речь идет о сделках – *a property deal* и *deal involving millions*); и, наконец, в оставшихся строках *deal*, перед которым идут прилагательные *good* или *great*, означает количество – *a good deal of concern*, а *great deal of money* и *working a good deal harder*.

С помощью компьютера достаточно легко получить полный список всех употреблений конкретного слова в контексте, а через контекст – все значения слова. Полный список конкордансов слова *deal* выдает много других значений, помимо перечисленных трех:

- *"to have a topic" (Two articles in astronomy deal in turn with stellar evolution and the determination of stellar distances);*
- *"a kind of treatment" (So let us see them get a fair and square deal);*
- *"a type of wood" (He went to the far end of the deal table and sat against it ...)*

Несмотря на то, что конкордансы, созданные из корпусов текстов, дают более солидную эмпирическую базу, чем традиционные лексикографические методы, иногда их данные становятся чрезмерными. Например, 8-миллионный корпус *Longman-Lancaster Corpus* предлагает 1500 конкордансов для слова *deal*. Простое выявление различных значений слова в такой огромной базе данных превратится в изнурительную работу, а точная группировка значений и их распределение по степени важности едва ли будет возможной без применения дополнительных инструментов.

5.2.2. Анализ частотности слова

Какие слова в языке являются самыми употребительные? Какие – самыми редкими? В какой части непрерывного континуума от наиболее до наименее употребительных слов будет нужное нам слово? Ответы на подобные вопросы являются первым шагом на пути понимания моделей употребления слова.

Например, частотный список для десяти выбранных текстов из корпуса *LOB*, созданный программой *TACT (Text-Analysis Computing Tools*), показывает, что слово *a* встречается 478 раз, тогда как слово *abandoned* только дважды. Данный перечень расположен в алфавитном порядке, хотя обычно частотные списки располагаются в порядке убывания по частотности. Аналогичный список можно получить для различных форм слова *deal* во всем корпусе *LOB*, из которого видно, что форма *deal* встречается 182 раза, *dealing* – 52 раза, *deals* – 25 раз и *dealt* – 31 раз. Обычно при анализе слова необходимо рассмотреть все его формы. Хотя компьютер воспринимает *deal, deals, dealing* и *dealt* как разные слова, может понадобиться рассмотреть их вместе или определить их общую частотность. Для обозначения базовой формы слова без учета его грамматических форм, таких как время или число, используется термин «лемма», таким образом, частотность леммы DEAL – *включая deal, deals, dealing* и *dealt* – в корпусе LOB составляет 290. Для разграничения между формами

слова и леммой последняя обычно обозначается малыми прописными буквами – DEAL.

Однако на первый взгляд трудно понять, что означает тот факт, что DEAL встречается в корпусе *LOB* 290 раз. Много это или мало? Если мы сравним лемму DEAL со служебными словами, например, артиклями или предлогами, то увидим, что она встречается относительно редко. Определенный артикль *the* встречается более 50 000 раз, а предлог *of* – более 35 000 раз. Если же мы сравним употребление DEAL с другими знаменательными словами, то увидим, что оно не является ни особенно частотным, ни особенно редким. Самым употребительным глаголом в корпусе является MAKE, который встречается 2417 раз, что гораздо больше чем DEAL. С другой стороны, глагол SIGH встречается только 16 раз.

Частотные списки корпуса общего языка показывают, что для анализа значения и употребления слов нужен очень большой корпус – корпус размером в 1 миллион слов не может дать достаточно данных для того, чтобы мы можно было делать обоснованные обобщения. Частотность является относительно надежным показателем в отношении преобладающего количества наиболее употребительных слов в корпусе, но для анализа значений и моделей сочетаемости требуется гораздо большее число употреблений. Например, опираясь на представленность леммы SIGH в корпусе *LOB*, можно лишь сказать, что оно является редким. Более того, в небольших корпусах включение малочастотных или редких слов во многом зависит от тематики текстов, представленных в корпусе. Так, например, в корпусе *LOB* слово *clowns* представлено только один раз, а слово *clown* – ни разу. Слово *elephant* встречается 10 раз, а *giraffe, zebra* или *hippopotamus* – ни одного. Если бы в этом корпусе был рассказ о клоуне или статья о жирафах, то эти цифры были бы совсем другими. Большие корпусы, включающие в себя тексты разнообразной тематики, меньше зависят от отдельных текстов.

Другой сложностью при анализе семантики слова является то, что одна и та же форма слова может выполнять различные функции. Например, *deals* может быть формой глагола в третьем лице

единственном числе и множественным числом существительного. Если слова в корпусе не имеют соответствующей разметки, то невозможно определить, какие грамматические формы являются употребительными, а какие – нет. Для того, чтобы разделить употребления слова *deals* в качестве существительного от глагола, исследователю придется просмотреть все формы в контексте, определить их грамматическую категорию и посчитать их. В случае с *deal* (182 употребления) в корпусе *LOB* это достаточно несложно, но для более частотного слова *look* потребуется гораздо более трудоемкая работой. Для повышения эффективности используются размеченные (аннотированные) корпусы. Так, слово *deal* употребляется 115 раз в качестве существительного в единственном числе (*nn*), один – в качестве имени собственного (*np*) и 66 – в качестве глагола (*vb*); форма *deals* встречается 5 раз как существительное во множественном числе (*nns*) и 20 раз – как глагол третьего лица единственного числа (*vbz*); форма *dealing* – один раз как существительное в единственном числе (*nn*) и 51 раз как причастие настоящего времени (*vbg*); и, наконец, *dealt* встречается 14 раз как глагол в форме простого прошедшего времени (*vbd*) и 17 раз – как причастие прошедшего времени (*vbn*). Имея такую информацию, можно более детально анализировать распределение глагольных и именных форм, а также сопоставлять их употребление в различных регистрах.

5.2.3. Распределение слова по регистрам

С помощью размеченного корпуса можно проследить распределение слова DEAL в качестве существительного и глагола по регистрам. Рассмотрим лемму DEAL в качестве существительного в корпусе *LOB*. Данный корпус состоит из различных регистров, таких как «научная проза» (*learned and scientific prose*), «беллетристика» (*belles lettres*) и «пресс-репортажи» (*press reportage*). Распределение частотности употребления данного существительного в единственном и

множественном числе в каждом из регистров может быть представлено в ненормированной и нормированной форме. Ненормированные цифры дают нам фактическое количество употреблений – например, в категории «научной прозы» – 16 раз. Однако данные категории имеют различное количество слов – пресс-репортаж – 88 тыс., а научная проза – 160 тыс., вследствие чего трудно утверждать, является ли слово более употребительным в одном регистре по сравнению с другим. Для решения этой проблемы используются нормированные подсчеты, преобразовывающие количество употреблений слова под стандартный масштаб, в данном случае – на 100 тыс слов. Теперь можно увидеть, что пресс-репортаж демонстрирует 15,9 употреблений на 100 тыс. слов, а научная проза – 10 на 100 тыс. слов. Нормированный подсчет помогает избежать ошибочных выводов и дает прочную основу для сопоставления словоупотреблений по регистрам.

Сопоставления в употреблении DEAL в качестве существительного и глагола в корпусе *Longman-Lancaster Corpus* для регистров художественной литературы (*fiction*) и академической прозы (*academic prose*) также дают интересные результаты. Общие нормированные подсчеты показывают, что глагол встречается немного чаще, чем существительное (119 слов на 1 млн к 90 словам на 1 млн). Однако, если обратиться к конкретным регистрам, то картина сильно изменяется. В академической прозе глагол встречается почти в два раза чаще, чем существительное (176 к 74 словоупотреблений на 1 млн слов), а в художественной литературе наоборот существительное представлено значительно больше, чем в глагол (107 к 63 словоупотреблений на 1 млн слов). Это указывает еще на одну важную вещь – корпус, ограниченный одним регистром, не может отражать функционирование языка в других регистрах.

5.2.4. Распределение значений слова по регистрам

Ранее говорилось о том, что корпус позволяет изучать значения слов через их использование в конкордансах, дающих полный перечень всех употреблений слова в контексте. Однако при этом возникает новая трудность – как можно рассортировать и проанализировать всю информацию конкордансов, если, например, для не самого частотного слова DEAL она содержит 2000 употреблений на 10-миллионный корпус? Вместо того, чтобы сортировать конкордансы вручную, можно начать анализ значений слов по их коллокатам – словам, с которыми чаще всего встречается данное слово.

Возьмем, например, лемму DEAL в качестве существительного в двух регистрах – академическая проза и художественная литература – в корпусе *Longman-Lancaster Corpus*. Наиболее частотным левым коллокатом будет слово *good*, а правым – *of*. Та же процедура может быть проделана для слов, отстоящих от заданного слова на две или три позиции. В академической прозе самой употребительной левой коллокацией будет слово *great* (45 раз на 1 млн слов), за которым идет *good* (23 на 1 млн). Вместе они дают 185 из 196 колокаций со значимой частотностью. Что это говорит нам о значении слова *deal*? Представляется вероятным, что *good/great deal* относится или к количеству чего-либо или к деловой сделке. Если мы обратимся к правой коллокации, то самым употребительным будет слово *of* (39 на 1 млн). Таким образом, можно предположить, что наиболее употребительным значением слова *deal* будет количество – *a good/great deal of*.

В художественной литературе это значение также является самым употребительным, но, в отличие от академической прозы, достаточно частотны и другие употребления. Например, когда левым коллокатом является артикль *the*, лемма DEAL имеет значение соглашения – *part of the deal is …* или *Isn't that the deal?* Коллокат *big* передает значение отсутствия важности – *no big deal* или *what's the big*

deal? Кроме того, многие коллокаты, которые не имеют высокой частотности по отдельности, вместе связаны с еще одним важным значением данного слова – *property deal, record deal, cash deal, land deal, mining deal*. Текст художественной литературы дает еще одно значение DEAL – в сочетании с правым коллокатом *table* и *box*, оно означает сорт дерева. В целом, можно сказать, что самое распространенное значение DEAL в качестве существительного – это указание на количество, однако художественная литература демонстрирует более высокую его частотность и большее разнообразие других значений, таких как «соглашение», «сорт дерева», «отсутствие важности».

Как эти данные соотносятся со словарными дефинициями? Обзор словарей показывает большое количество значений слова DEAL:

1) *a large but indefinite amount;*
2) *an agreement or arrangement;*
3) *the distribution of cards in a game;*
4) *treatment received;*
5) *the act of distributing;*
6) *wood of fir or pine trees;*
7) *the act of buying or selling or a business transaction.*

Однако и среди этих основных определений существует большое разнообразие. В большинстве словарей представлены все семь значений, но в двух отсутствует значение "*act of distributing*", а в одном – "*agreement/arrangement"*. Порядок дефиниций также значительно варьируется. Так, "*a large but indefinite amount*" является вторым значением в первой статье словаря *Webster's Third Dictionary,* но только двадцать первым значением в словаре *Random House Dictionary*.

Таблица 6. Распространенные словарные дефиниции слова DEAL в качестве существительного (Biber 2004)

Определение	*Webster's Encyclopedic Dictionary*	*Webster's Third Dictionary*	*Chambers Dictionary*	*Random House Dictionary*	*Longman Dictionary of English Language and Culture*
	1989	***1981***	***1993***	***1993***	***1992***
large but indefinite amount	Статья 1 значение 16	Статья 1 значение 2	Статья 1 значение 3	Статья 1 значение 21	Статья 2 значение 3
agreement/ arrangement	Статья 1 значение 13	Статья 3 значение 3	_	Статья 1 значение 18	Статья 2 значение 1
distribution of cards in a game	Статья 1 значение 18	Статья 1 значение 3	Статья 1 значение 4	Статья 1 значение 21	Статья 2 значение 4
treatment received	Статья 1 значение 15	Статья 3 значение 2	Статья 1 значение 6	Статья 1 значение 6	Статья 2 значение 2
act of distributing	Статья 1 значение 17	_	_	Статья 1 значение 23	_
pine or fir wood	Статья 2 3 значения	Статья 4 2 значения	Статья 2 значение 1	Статья 2 3 значения	Статья 3 значение 1
act of buying or selling/ a business transaction	Статья 1 значение 13	Статья 3 значение 2	Статья 1 значение 5	Статья 1 значение 17	Статья 2 значение 1

Сопоставление данных дефиниций с результатами корпусных исследований значений слова DEAL ставит ряд вопросов:

- самым частотным значением в корпусе является «количество», но в словарях оно является далеко не первым – в двух случаях оно является шестнадцатым и двадцать первым;
- анализ коллокатов обнаружил наличие достаточного частотного значения, не нашедшего отражения в словарях – *big deal* как выражение отсутствия важности;
- наконец, ни один словарь не отражает различий по регистрам, хотя более современные словари, основанные на корпусах, уже делают это;
- все пять словарей дают значение, которое ни разу не встречается в корпусе – «раздача карт в игре». Хотя для

большинства носителей языка это первое значение, с которым ассоциируется слово *deal*, в реальном языке оно встречается редко. Исключением является достаточно ограниченная сфера карточных игр.

Поэтому современные словари, основанные на корпусах, начиная с *Macmillan English Dictionary* и *Longman Dictionary of Contemporary English*, стремятся охватить все значения и оттенки значения слов.

5.2.5. Анализ синонимов

Подход к анализу близких по значению слов в корпусной лингвистике основан на выявлении систематических отличий в моделях употребления синонимичных слов. Возьмем, к примеру, такие синонимичные прилагательные, выбор между которыми очень часто вызывает затруднения у изучающих английский язык, как *big, large* и *great*. Все они означают размер. Проанализируем их употребление на предмет выявления отличий.

5,7-миллионая выборка из корпуса *Longman-Lancaster Corpus* без учета различий регистров дает следующую частотность данных прилагательных: самое употребительное – *large* (нормированная частота 408 на 1 млн), далее идет *great* (393) и затем – *big* (230).

Для академической прозы последовательность та же, но данные по частотности – иные. Так, *large* является самым употребительным (605/1 млн), а *big* встречается намного реже – только 31/1 млн. В художественной литературе *great* и *big* имеют похожие показатели (соответственно 490 и 408), тогда как *large* имеет более низкую частотность (232).

Таким образом, употребление этих, казалось бы, синонимичных прилагательных сильно отличается от регистра к регистру. *Big* в художественной литературе встречается более чем в 10 раз чаще, чем в академической прозе, тогда как *great* в полтора раза чаще встречается в художественной литературе. С другой стороны, *large* в академической прозе встречается в три раза чаще, чем в художественной литературе. Для того, чтобы понять причины таких отличий, следует обратиться к коллокатам данных прилагательных.

В обоих регистрах коллокаты прилагательного *big* свидетельствуют о том, что чаще всего оно относится к физическому размеру. В академической прозе употребительными являются коллокаты *big enough* (6/2,2 млн слов) и *big traders* (3/2,2), причем, последний представлен одним текстом об экономическом развитии в Западной Африке, где сопоставляются крупные и мелкие трейдеры.

В художественной литературе гораздо больше коллокатов, где данное прилагательное относится к размеру физических объектов, таких *man, house, toe, boy, room* и неопределенное местоимение *one*. Кроме того, *big* употребляется с другими определениями, такими как *black, old* и *red* (*… and big black eyes*). Аналогичным образом, сочетание *big enough* в художественной литературе употребляется для указания на физический размер: *The cart was not really* ***big enough***. И, наконец, распространенным коллокатом прилагательного *big* является союз *and*. Однако необходимо отметить его частую сочетаемость со всеми тремя прилагательными.

В отличие от *big* коллокаты прилагательного *large* в академической прозе показывают, что оно чаще всего употребляется для обозначения количества чего-либо. Первые семь коллокатов имеют четко выраженное значение количества: *large + number(s), proportion, amount(s), quantities, part, extent*. Сочетание *large+enough* также часто указывает на количество или пропорции: *The ratio is* ***large enough****, however, to allow…* . Это же сочетание употребляется для указания на физический размер. И, наконец, употребительным в академической прозе является сочетание *large scale*, которое используется как устойчивое неразделяемое определение для указания на величину различных процессов: *large-scale centralisation.*

В художественной литературе прилагательное *large* чаще всего употребляется в значении «физический размер» с существительными и прилагательными: *house, room, man, black и white: a large white bird.*

Данное употребление *large* идентично наиболее частотному употреблению *big* в художественной литературе, и многие правые коллокаты встречаются с обоими прилагательными. Однако в целом *large* реже используется в художественной литературе и,

следовательно, эти словосочетания также встречаются реже. Например, *big man* встречается 9,6 раз на 1 млн, тогда как *large man* только два раза на 1 млн, *big house* – 7,6/1 млн, *large house* – 3/1 млн. *Large* также используется для указания на количество. Сочетание *large number* является сравнительно частотным в художественной литературе. Другие правые коллокаты, такие как *amount, proportion* и *sum,* по отдельности встречаются редко, но вместе образуют класс словосочетани, указывающих на количество: *A* ***large number*** *of people sat round a table ….*

Третье прилагательное, *great*, имеет свою коллокационную модель. В академической прозе его самое частотное значение – указание на количество, и самым употребительным (с большим отрывом) коллокатом является *great deal*. Однако кроме него также встречаются *great number, great majority, great variety, great extent* и *great part*. Это употребление прилагательного *great* в академической прозе идентично употреблению *large*, хотя large никогда не сочетается с *deal*. *Great* имеет еще одно значение – указание на интенсивность. В академической прозе это употребление встречается с такими правыми коллокатами как *importance, care, advantage, detail* и *interest*: *The figures have to be interpreted with* ***great care***.

В художественной литературе *great* чаще всего употребляется для указания на количество, и самым частотным сочетанием является *great deal*: *He stood and drank a* ***great deal*** *of apple juice*. Однако в отличие от академической прозы в художественной литературе *great* имеет гораздо больше значений. Например, сочетание *great man* встречается 6,6 раз на 1 млн слов в значении «очень важный» или «хороший». В сочетании *great care* прилагательное *great* также означает «очень хороший»: *… I promise you we will take* ***great care*** *of him*. Кроме того, *great* иногда используется для указания на большой физический размер: *a great, black bird.*

В художественной литературе *great* может также употребляться в качестве интенсификатора для прилагательного *big: it's a* ***great big*** *country with a continent of promise*. Значение интенсивности идентично тому, что встречается в академической прозе: *great care, great pleasure*

и *great relief*. И, наконец, *great* имеет еще одно специфическое употребление – указание на родство: *great aunt* (сестра бабушки)*: He was almost as old as her* ***great aunt*** *had been*.

Таким образом, мы видим, что несмотря на большое разнообразие, в употреблении данных трех синонимов выделяются определенные модели. *Big* чаще всего указывает на физический размер, тогда как *large* употребляется для обозначения количества. *Great* также означает количество, особенно в сочетании *great deal*, но имеет большее количество значений – от интенсивности до размера и родства.

Распространенные значения у трех прилагательных помогают объяснить их различную частотность в двух регистрах. Тексты художественной литературы содержат больше описаний физических объектов, относящихся к размеру предметов, людей, мест. Когда речь идет о размере в академической прозе, то здесь чаще используются конкретные единицы измерения. Поэтому, логично, что *big* чаще встречается в художественной литературе, чем в академической прозе. Последняя больше относится к количеству, и поэтому здесь чаще встречается прилагательное *large*. В обоих регистрах *great* употребляется для указания на количество с помощью сочетания *great deal*. Однако в художественной литературе используется больше значений прилагательного *great*, что связано с более разнообразными описаниями и темами по сравнению с академической прозой.

Слова имеют ассоциативные связи не только со своими ближайшими соседями. Два слова могут часто встречаться вместе, даже если между ними находятся несколько других слов. Рассмотрим второй правый коллокат прилагательного *large*, т.е. *of* в сочетании *large X of*. Чаще всего это сочетание встречается в академической прозе – настолько, что его можно рассматривать как своего рода фрейм, в который можно подставлять различные слова: *large amount of, large proportion of* и *large group of*. Как отмечалось ранее, прилагательное *large* чаще всего встречается с существительными, указывающими на размер. Анализ такого расширенного фрейма

показывает, что эти словосочетания выполняют две основные функции:

1) указание на размер или меру измерения: *a large number of, large numbers of, large amounts of, large quantities of, a large volume of, large bundles of, large masses of, a large batch of, large areas of;*
2) указание на часть большего: *a large proportion of, a large part of, a large sample of, a large fraction of, a large segment of.*

В художественной литературе интерес представляет модель *large X eyes*. Оно встречается гораздо чаще (4,3/1 млн), чем просто *large eyes* (1,6). Это объясняется тем, что существительное *eyes* помимо определения *large* часто определяется прилагательными, обозначающими цвет или качество в следующей последовательности: *large* + прилагательное, означающее цвет/качество + *eyes*. Например:

his large hazel eyes
large brown eyes
large black eyes
very large dark eyes
large watery eyes.

Корпусный анализ помогает разграничить синонимы по различиям в их употреблении, модели которых являются весьма систематичными и достаточно сложными.

§ 5.3. Корпусные жанровые исследования

Легкость доступа к огромным массивам разнообразного лингвистического материала при помощи все более доступного компьютера и стандартного и простого в использовании программного обеспечения привело к принципиально новым результатам, когда пользоваться корпусом стало доступно широким кругам лиц, занимающимся изучением языка.

Примером такого проекта может служить корпус англоязычной деловой корреспонденции VOBEC, созданный на кафедре лингвистики и межкультурной коммуникации Поволжского филиала Международного университета в Москве (Толстова 2007). В качестве основной задачи проекта было поставлено формирование электронного корпуса англоязычной корреспонденции, состоящего из следующих основных массивов:

1) письма-образцы, заимствованные из справочных изданий авторитетных британских и американских авторов, предназначенных для бизнесменов – носителей английского языка, ведущих коммерческую переписку;

2) аутентичные деловые письма и электронные сообщения на английском языке, написанные носителями языка (из США, Великобритании, Канады, Австралии, Ирландии);

3) деловые письма и электронные сообщения, написанные на английском языке лицами, не являющимися носителями данного языка (гражданами России, ближнего и дальнего зарубежья). Такого рода материал необходим для проведения сопоставительного языкового анализа текстов носителей и не-носителей английского языка (лиц, для которых он является иностранным);

4) деловые письма и электронные сообщения, написанные на английском языке лицами, для которых не установлено, являются они носителями данного языка или нет.

Общий объем корпуса составил в итоге 769 файлов (1 файл = 1 письмо / факсимильное сообщение / электронное сообщение), состоящих из 80.347 словоупотреблений (*tokens*), представленных 6.757 словами (*types*). Из общего количества слов более трети (2.571)

составляют так называемые гапакс легомена (лат. *hapax legomena*), т.е. слова, встречающиеся в корпусе лишь один раз.

Рассмотрим для примера такой распространенный жанр деловых писем, как письмо-извинение. Данный жанр тесно связан с категорией вежливости, от владения которой может зависеть успех или, наоборот, неуспех деловых отношений. Всего в корпусе таких писем было выделено 135, критерием отнесения писем к данному жанру было наличие в них таких формальных лексических средств выражения извинения, как *sorry, regret, apologise* и их производных.

С помощью инструмента *Wordlist* из программы *WordsmithTools* было выявлено, что наиболее частотными выражениями извинения являются следующие:

- *sorry* – 56 писем (42%);
- *regret* (31 письмо) и его производные, такие как *regretfully* (3 письма), *regrettable* (1 письмо) – всего 36 писем (27%);
- *apology* (5 писем) и его производные, такие как *apologise/apologize* (21 писмо), *apologies* (16 писем) – всего 42 письма (31%).

Ниже приводятся фрагменты конкордансов для слова *regret*.

Рис. 11. Фрагмент списка конкордансов для прилагательного *regret*

Следует отметить, что *sorry* является самым неформальным средством выражения извинения, тогда как *apologise* – наиболее формальным. Промежуточное положение занимает глагол *regret*, который может звучать формально и нейтрально.

Наиболее частотными сочетаниями с данными единицами являются следущие:

1. с глаголом *to regret*:
- *I regret I cannot give you an immediate order.*
- *We regret that we cannot accept this payment as a full discharge.*
2. С глаголом *to apologise*:
- *We apologise for any inconvenience this may cause and look forward to hearing from you shortly.*

а также однокоренным существительным *apology* в единственном и множественном числе:

- *Again, please accept my personal apology for any inconvenience.*
- *Please accept my apologies for the late reply – we are extremely busy at the moment.*

3. С прилагательным *sorry*:
- *Dear Tina, sorry for causing more trouble.*
- *We are sorry to hear that you are experiencing problems receiving your magazines.*

Письма-извинения не ограничиваются только формулировкой самого извинения – нормы вежливости требуют объяснить причину нарушения взятых обязательств или допущенных ошибок. Для этих целей используются различные грамматические конструкции: страдательный залог, модальные глаголы, видовременные формы глагола и т.п.

Страдательный залог употребляется в качестве средства отдаления, снятия ответственности за неприятное событие. Обращает на себя внимание тот факт, что в письмах-извинениях содержащих лексему *apology* – наиболее формальное выражение вежливости – страдательный залог употребляется чаще, чем в остальных письмах-извинениях, напр.:

Dear Mr Daniels

I am sorry to learn from your letter of 23 August that you find our prices too high. We do our best to keep prices as low as possible without sacrificing quality. We are sorry to inform you that high income tax ***was taken*** *for the shipment of your order. Unfortunately, it does not depend on our company.*

С помощью страдательного залога автор письма снимает с себя ответственность за неприятную ситуацию, не обозначая виновника ошибки.

Модальные глаголы *to be able/to be unable* также играют важную роль в письмах-извинениях. Основной функцией данных глаголов, выражающих невозможность выполнить действие, является объяснение причин (*Giving reasons*), невыполнения обязательств, задержек с поставками и т.п.:

– *We have not been able to identify your subscription from the information provided.*
– *We are unable to go any higher than 7 %.*

Еще одним важным грамматическим средством в письмах-извинениях является употребление будущего времени (*Future Simple*). Основной функцией глагольной формы будущего времени *Future Simple* является обещание принять меры по устранению неисправностей и недопущению их повторения в будущем (*Actions to be taken*). При этом форма *will* для 1-го лица указывает на дополнительное модальное значение:

– *We apologise for any inconvenience and we will send you the order immediately.*

Что касается риторической структуры писем-извинений, то в них прослеживаются регулярно повторяющиеся элементы. Обращает на себя внимание, что сам акт извинения реализуется дважды – в начале и в конце письма. Первое извинение включает в себя объяснение причины ошибки. Далее следует указание на то, какие меры были приняты или будут приняты в будущем. Вторичное извинение является обращением к получателю, извинением за причиненное неудобство, но оно, как правило, является более кратким:

– *Please accept our apologies once again.*

– *We hope that this has not caused you any inconvenience.*

– *With apologies once again.*

В результате анализа писем–извинений, содержащихся в корпусе, была выявлена четкая риторическая структура, включающая в себя четыре основных коммуникативных шага (*moves*):

Move 1 – извинение, содержащее главную информацию (*The First Apology*).

Move 2 – объяснение причины произошедшей ошибки/недоразумения (*Giving Reasons*).

Move 3 – информация о принятых/будучи принятых мерах (*Informing of Actions taken/to be taken*).

Move 4 – вторичное (окончательное) извинение (*The second Apology*).

Ниже приведен пример письма-извинения, построенного в соответствии с данной структурой:

Dear Mr Taylor

Move 1:

We are sorry to learn from your letter of 10 May of the difficulties you are having with the pens supplied to your order number 8562.

Move 2:

All our pens are manufactured to be identical in design and performance and we cannot understand why some of them should have given trouble to your customers. It is normal practice for each pen to be individually examined by our Inspection Department before being passed into store. However, from what you say, it would seem that a number of the pens included in the latest batch escaped the usual examination.

Move 3:

We sympathise with your problem but regret that we cannot accept your suggestion to take back all the unsold stock from the batch concerned. Indeed there should be no need for this since it is unlikely that the number of faulty pens can be very large. We will gladly replace any pen found to be unsatisfactory, and on this particular batch are prepared to allow you a special discount of 5% to compensate for your inconvenience.

Move 4:

We trust you will accept this as being a fair and reasonable solution of this matter.With my apologise once again,

Yours sincerely

Ниже приведен график дисперсии для единицы *apologise*, показывающий, в какой части текста встречается то или иное слово (в данном случае это последняя треть текста).

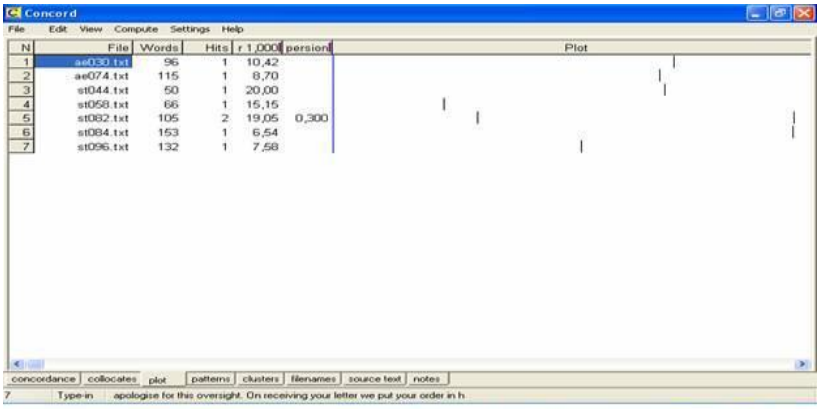


Рис. 12. График дисперсии слова *apologize* в корпусе писем-извинений

Данное мини-исследование показывает, что при выборе языковых средств в процессе речепроизводства ключевыми являются два момента: учет жанровых особенностей и требований идиоматичности.

§ 5.4. Использование корпуса при обучении иностранному языку

Возможность применения корпуса не только в исследовательских целях, но и в прикладных, практических целях преподавания английского языка, позволяет сделать обучение более осознанным, интересным, а потому – эффективным.

Одним из достоинств корпусной лингвистики является освобождение от «монополии» интуиции носителей языка. Использование аутентичного материала корпусов позволяет исследователям-инофонам выдвигать свои собственные предположения и гипотезы, находя для них достоверные обоснования, самостоятельно делать выводы и обобщения, не обращаясь к знаниям и «языковому чутью» лиц, для которых английский является родным языком (Simpson, Swales 2001).

И, наконец, следует отметить возможность применения корпусной лингвистики не только в исследовательских целях, но и для преподавания английского языка, делая при этом обучение более осознанным, интересным, а потому – эффективным.

До сих пор корпусная лингвистика большей частью рассматривалась только в контексте лингвистической науки. Тем не менее, непростительной ошибкой было бы пройти мимо тех возможностей, которые она способна дать в плане обучения иностранному языку. Использование корпуса позволяет задействовать одновременно несколько современных подходов к обучению.

Гуманистический подход к обучению состоит в ориентации на личность обучаемого, на его интересы, потребности и возможности, а также на способы учения, которые соответствуют его индивидуальным особенностям. Для данного подхода характерна переориентация всего процесса обучения с личности преподавателя и методов преподавания на личность учащегося и способы учения. Особенность обучения с помощью корпуса как раз и состоит в том, что здесь преподаватель не ставит перед собой задачу научить

студента – напротив, осуществляя поиск самостоятельно, тот учится узнавать новое.

С точки зрения общей направленности процесса мышления можно выделить дедуктивный и индуктивный подходы. Традиционно в преподавании иностранных языков применялся дедуктивный подход, который предусматривает объяснение правила и его тренировку на практике, т.е. путь от общего к частному, от формы к ее реализации. Индуктивный подход, напротив, предполагает путь от частного к общему, от функционирования того или иного грамматического или лексического явления в разнообразных контекстах– к пониманию его значения и формы.

Индуктивный подход к обучению в современной методике привел к появлению нового термина *consciousness-raising approach* – сознательно-ориентированный подход, который получил широкое распространение в **компьютерном обучении** (*computer assisted language learning*) – еще одном инновационном подходе, который, в свою очередь, тесно связан с **обучением с помощью баз данных** (*data-driven learning*) – подходом, опирающимся на индуктивные процессы познания и языковые базы данных, содержащие аутентичные устные и письменные высказывания. При этом используемый корпус языковых данных выступает в качестве материала, а соответствующее программное обеспечение – инструмента обучения.

Основная цель данного подхода – научить учащихся самостоятельно извлекать из аутентичного материала различного рода лингвистическую информацию об особенностях его употребления. Студенты, как правило, работают индивидуально в компьютерном классе, используют специальные компьютерные программы и без помощи преподавателя анализируют интересующие их лингвистические явления. Согласно этому подходу учащемуся следует развивать умение учиться, он должен выступать в роли исследователя, самостоятельно решать задачи, связанные с осознанием языковой формы.

Рис. 13. Обучение с помощью корпуса
в системе современных подходов к обучению

Такая самостоятельная работа представляет большой интерес для учащихся, «бросая вызов» их возможностям в работе и с языковыми явлениями, и с компьютером. Учитывая тот факт, что зачастую многие студенты знают компьютер лучше, чем некоторые преподаватели иностранного языка, это дает учащимся возможность почувствовать себя «на равных» с преподавателем, в отличие от привычной иерархии «учитель-ученик» или «преподаватель-студент». Очевидно, что наибольшую пользу из работы с корпусом извлекут студенты, увлекающиеся компьютерами. Здесь можно говорить о «технофилах» и «технофобах» (Dudley-Evans, St John 2003): для первых новая технология представляет дополнительный интерес и стимул, в то время как на последних (которых, к счастью, все-таки меньше) она не окажет столь значительного влияния.

Обучение иностранному языку должно ставить перед собой, прежде всего, практические задачи, достижение которых осуществляет движение вперед. Среди основных целей можно выделить приобретение учащимися конкретных навыков и умений, которые они в дальнейшем смогут применять в своей деятельности (Щукин 2004). К сожалению, нередко возникает перекос в сторону рецептивных навыков (письменных – в форме чтения – и устных – в форме аудирования) за счет продуктивных – производства устной и письменной речи, уместной в той или иной ситуации профессиональной деятельности. Чтобы приобрести эти навыки, нельзя обойтись без работы с корпусом соответствующего аутентичного языкового материала. Такая работа поможет сделать речь (как устную, так и письменную) идиоматичной, то есть приближенной к речи носителей языка.

Это можно в полной мере отнести к такой области английского языка как язык делового общения. Только всесторонне изучив и проанализировав его различные регистры и жанры, можно подойти к самостоятельному составлению писем и другой документации, проведению презентаций, переговоров и т.п. (Назарова 2004).

При составлении собственных текстов деловых писем студенты смогут легко и быстро находить в корпусе подходящие слова и сочетания слов, соответствующие тематике и стилю каждого вида писем, а также грамматические формы и конструкции, используемые для реализации той или иной коммуникативной функции. Это поможет им избежать распространенной тенденции пытаться сочинять письмо на русском языке, а затем переводить его на английский. Для преподавателей же корпус представляет собой неисчерпаемый источник создания собственных творческих заданий и упражнений на полностью аутентичном материале.

Рассмотрим конкретные возможности, которые предлагает нам компьютеризированный корпус для обучения английскому языку делового общения. Возьмем для начала инструмент определения сочетаемости слов, конкордансер, на примере слова *advise*.

Во-первых, **множественный контекст** (*multiple context*) позволяет выявить различные значения и оттенки значения слова.

Таблица 7. Множественный контекст на основе списка конкордансов слова *advise*

Please	**advise**	as to repair or replacement.
Please	**advise**	if this is not possible.
Please	**advise**	me of your decision.
anything that I need to prepare, please	**advise**	me.
However, we regret to	**advise**	that we concluded to decline your
investigate how much will be needed and	**advise**	us as soon as possible.
ill appreciate it very much if you will	**advise**	us of a date convenient to you.
Please also	**advise**	us what we are to do with the rugs
Should this plan change, we will	**advise**	you immediately.
Their consultant will be able to	**advise**	you on what preparation is necessary
We are pleased to	**advise**	you that thanks to their kind assistanc
We therefore	**advise**	you to make alternative arrangement

Студентам предлагаются многочисленные примеры употребления лексических единиц в контексте. Они анализируют эти однострочные конкордансы, отмечая окружение ключевых слов и пытаясь по контексту определить их употребление. Затем им предлагается выполнить разнообразные виды деятельности, напр.:

- внимательно рассмотреть конкордансы ключевого слова и окружающие его слова и постараться определить его значения;
- ознакомиться с речевыми образцами, окружающими ключевое слово, обращаясь к конкордансам, выполняя задания;
- потренироваться в употреблении ключевых слов, не обращаясь к конкордансам;
- создать свой собственный текст с ключевыми словами в соответствии с конкретным жанром делового письма.

Анализ различных значений ключевого слова, например, для конкордансов из Таблицы 7, можно предложить сделать следующим образом:

Какие из нижеприведенных предположений, являются, по-вашему, верными? Отметить правильный вариант галочкой.

TO ADVISE involves giving opinion	true ❑	false ❑
TO ADVISE involves offering an idea	true ❑	false ❑
TO ADVISE involves deciding	true ❑	false ❑
TO ADVISE involves informing someone	true ❑	false ❑
TO ADVISE involves acting as a professional adviser	true ❑	false ❑
TO ADVISE involves letting somebody know	true ❑	false ❑
TO ADVISE involves advertising something	true ❑	false ❑

Для отработки предлогов можно предложить сделать следующее упражнение:

Какие предлоги наиболее часто употребляются для каждого из значений слова to ***advise****?*

1. advise ……
2. advise ……
3. advise ……
4. advise ……

или:

Какие предлоги чаще всего встречаются с разными формами слова:

1. advise ……..
2. advised ……
3. advising ……...
4. advice ……..

Можно проанализировать наиболее употребительные формы слова. Напр., *advising* встречается только в роли причастия в функции определения и преобладающей моделью является *advising + object + preposition* (чаще всего – *of*), причем доминирующим лексическим значением у этого глагола в форме причастия будет «советовать», тогда как в личных глагольных формах первым значением (с большим отрывом) является «информировать, сообщать», что несколько

противоречит той последовательности, которая представлена в словарях общей лексики, где сначала дается значение «советовать», а лишь за тем – «информировать, сообщать», к тому же с пометой *'formal'* («формальный»).

Для отработки речевых образцов можно также использовать упражнения на заполнение пропусков (*gapping*) и соотнесение единиц языка (*matching*), напр.:

В каждой группе конкордансов одно из изученных вами слов пропущено. Определите, какое слово было удалено:

and I hope you can give me some valued	about it.
We took your	and added one two months ago.
Without your	and constant attention to detail, we wo
We seem to have ignored BW's	and potentially wasted BCC maintenance
ancy in terms between *****Bank's cable	and the actual guarantee now in our hom
received by our bank, and we have your	of dispatch.
It's good	

Для определения соотнесения единиц языка в упражнении может быть задание провести соединительные линии между предложениями в левом столбике и подходящими к ним по смыслу предложениям из правого столбика.

Жанры деловых писем (просьба, запрос, предложение, ответ на просьбу и т.п.) также могут послужить основой создания разнообразных заданий и упражнений, таких как поиск речевых моделей, наиболее часто встречающихся в конкретном жанре, а также анализ различных **коммуникативных ходов** (*moves*), из которых построены тексты рассматриваемого жанра. Студентам можно предложить провести такое мини-исследование:

В различных культурах письма-просьбы имеют различную структуру с точки зрения того, в какой его части лучше всего поместить саму просьбу. Например, в Китае просьба ставится обычно в конце письма и характерна следующая схема: приветствие, преамбула, причины и затем – просьба. В других культурах порядок может меняться, однако, преобладают два варианта: (1) Просьба + Причины / обоснования; (2) Причины + Просьба. Проанализируйте

письма-просьбы из корпуса и определите, какой тип в нем представлен.

Мы не рассматривали других возможностей корпуса при обучении английскому языку, например, использование электронного гипертекста непосредственно из сети Интернет, ресурсы для сопоставления текстов по сходной тематике на разных языках, что облегчило бы поиск переводческих эквивалентов при обучении переводу, и многие другие. Вне всякого сомнения он не может (и не должен) быть доминирующим методом, поскольку требует от учащихся достаточно много времени и сил, но нельзя отрицать тот факт, что в современных условиях информационной революции, с современными студентами и техническими средствами – корпусная лингвистика может существенно обогатить как лингвистов-исследователей, так и преподавателей принципиально новыми и эффективными методами работы, способствуя тем самым процессу оптимизации функционирования языка, которая предполагает выбор лучшего (оптимального) варианта из множества подобных.

Список литературы:

1. Агапова, С.Г. Основы межличностной и межкультурной коммуникации (английский язык). – Ростов н/Д.: Феникс, 2004.
2. Англо-русский словарь по лингвистике и семиотике / под ред. А.Н. Баранова и Д.О. Добровольского. – М.: Азбуковник, 2001.
3. Антология речевых жанров: повседневная коммуникация / под ред. проф. К.Ф. Седова. – М.: Лабиринт, 2007.
4. Арутюнова, Н.Д. Язык и мир человека. – М.: Языки русской культуры, 1998.
5. Арутюнова, Н.Д. Дискурс // Лингвистический энциклопедический словарь / гл. ред. В.Н. Ярцева. – М., 1990.
6. Ахманова, О.С. др. О точных методах исследования языка (О так называемой «математической лингвистике»). – М.: МГУ, 1961.
7. Базарова, Б.Б. Введение в корпусную лингвистику. – Улан-Удэ: Издательство Бурятского госуниверситета, 2016.
8. Баранов, А.Н. Введение в прикладную лингвистику. – М.: УРСС Эдиториал, 2001.
9. Барнет, В. Проблемы изучения жанров устной научной речи // Современная русская устная научная речь. – Красноярск: Издательство Красноярского университета, 1985. Т. 1.
10. Бахтин, М.М. Литературно-критические статьи. – М.: Художественная литература, 1986.
11. Бахтин, М.М. Проблема речевых жанров // Бахтин М.М. Собрание сочинений в 7 томах. – М.: Русские словари, 1996.
12. Вежбицка, А. Речевые жанры // Жанры речи. – Саратов: Колледж, 1997. Вып. 1.
13. Вежбицка, А. Речевые жанры [в свете теории элементарных речевых единиц] // Антология речевых жанров: повседневная коммуникация / под ред. проф. К.Ф. Седова. – М.: Лабиринт, 2007. – С. 68-80.
14. Всеволодова, А.В. Компьютерная обработка лингвистических данных. – М.: Флинта-Наука, 2007.
15. Вышкин, Е.Г. К проблеме соотношения терминов «текст» и «дискурс» в современном языкознании // Интеграция международных интеллектуальных процессов в образовании, бизнесе и межкультурной коммуникации. Сборник материалов международной научной конференции. 20-21 марта 2007 г. – Самара, 2007. – С. 147-150.

16.Гайда, С. Жанры разговорных высказываний // Жанры речи. – Саратов: Колледж, 1997. Вып. 1.

17.Гальперин, И.Р. Текст как объект лингвистического исследования. – М.: Едиториал УРСС, 2004.

18.Гаспаров, Б.М. Язык, память, образ. Лингвистика языкового существования. – М.: Нов. лит. обозрение, 1996.

19.Гаспаров, М.Л. Избранные статьи. – М.: НЛО, 1995.

20.Гвишиани, Н.Б. Практикум по корпусной лингвистике. – М.: Высшая школа, 2008.

21.Гвишиани, Н.Б. Терминология английского корпусного дискурса: метаязыковые различия и инновации // Филологические науки. Вопросы теории и практики. – 2016. – Т. 7, № 61. – С. 72–78.

22.Гвишиани, Н.Б., Герви, О.Ю. Корпусная лингвистика и грамматика речи // Вестник Московского университета. Сер. 9. Филология. 2001. № 2

23.Гвишиани, Н.Т. Корпусная лингвистика в изучении английского языка // Вестник МГУ. Серия 9. Филология. 1997. № 1.

24.Гольдин, В.Е. Теоретические проблемы коммуникативной диалектологии: дисс. в виде науч. докл. … докт. филол. наук. – Саратов, 1997.

25.Грайс, Г.П. Логика и речевое общение // Новое в зарубежной лингвистике (Лингвистическая прагматика) – М.: Прогресс, 1985.

26.Грудева, Е.В. Корпусная лингвистика. – М.: ФЛИНТА, 2012.

27.Гудков, Д.Б. Теория и практика межкультурной коммуникации. – М.: Гнозис, 2003.

28.Девкин, В.Д. Занимательная лексикология. – М.: ВЛАДОС, 1998.

29.Дейк, ван Т.А. Язык, познание, коммуникация. – М.: БГК им. И.А. Бодуэна де Куртенэ, 1999.

30.Дементьев, В. В. Теория речевых жанров. – М.: Знак, 2010.

31.Дементьев, В.В. Изучение речевых жанров: Обзор работ в современной русистике // Вопросы языкознания. 1997. № 1. – с. 109–121.

32.Дементьев, В.В. Непрямая коммуникация. – М.: Гнозис, 2006.

33.Дементьев, В.В. Фатические и информативные коммуникативные замыслы и коммуникативные интенции: проблемы коммуникативной компетенции и типология речевых жанров // Жанры речи. Саратов: Колледж, 1997. Вып. 1.

34.Захаров, В.П. Информационно-поисковые системы. – СПб.: Издательство СПбГУ, 2005.

35.Захаров, В.П. Корпусная лингвистика в россии // III Междисциплинарный научный форум Когниция. Коммуникация. Культура. CrossLingua'2014. Симферополь-Алушта, 2014. [Электронный ресурс]. – Режим доступа: http://crosslingua.cfuv.ru/publications/program2014.pdf. Проверено 01.10.18.

36.Захаров, В.П. Корпусная лингвистика. – СПб.: Издательство СПбГУ, 2006.

37.Зубов, А.В. Информационные технологии в лингвистике. – М.: Академия, 2004.

38.Карасик, В.Н. Языковой круг: личность, концепты, дискурс. – М.: Гнозис, 2004.

39.Кибрик, А.А., Плунгян В.А. Функционализм // Современная американская лингвистика: фундаментальные направления / под ред. А.А. Кибрика, И.М. Кобозевой и И.А. Секериной. – М.: Едиториал УРСС, 2002. – С. 276 – 339.

40.Китайгородская, М.В., Розанова Н.В. Речь москвичей: Коммуникативно-культурологический аспект. – М.: Русские словари, 1999.

41.Клюев, Е.В. Речевая коммуникация. – М.: ПРИОР, 1998.

42.Колесникова, Н.Л. Business Communication. – М.: Флинта: Наука, 2005.

43.Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 31 мая – 3 июня 2017 г.). Вып. 16 (23): В 2 т. Т. 1 – М.: Изд-во РГГУ, 2017.

44.Копотев, М. Введение в корпусную лингвистику. – Прага: Animedia Company, 2014.

45.Костомаров, В.Г., Бурвикова, Н.Д. Изучение и преподавание русского языка от слова Пушкина до наших дней // Материалы конференций и семинаров. – Волгоград: Изд-во Волгогр. гос. ун-та, 1999. – С. 12 – 18.

46.Кронгауз, М.А. Семантика. – М.: Academia, 2005. – С. 150–151.

47.Кубрякова, Е.С. Категоризация мира: пространство и время (Вступительное слово). Материалы научной конференции

«Категоризация мира: пространство и время». – М.: «Диалог-МГУ», 1997. – С. 3-14.

48.Куланов, М.Н. Управление кадрами: в помощь начинающему руководителю. – М.: Дашков и К°, 2005.

49.Левицкий, Ю.А. Лингвистика текста. – М.: Высшая школа, 2006.

50.Левицкий, Ю.А. Проблема типологии текстов. – Пермь: Изд-во Перм. ун-та, 1998.

51.Лихачев, Д.С. Концептосфера русского языка. – СПб.: РАН-СЛЯ, 1993.

52.Макаров, М.Л. Интерпретативный анализ дискурса в малой группе. – Тверь: Изд-во Твер. ун-та, 1998.

53.Макаров, М.Л. Основы теории дискурса. – М.: Гнозис, 2003.

54.Маслова, В.А. Лингвокультурология. – М.: Академия, 2001.

55.Назарова, Т.Б. Когнитивный подход и корпусные исследования: соперничество, союз или слияние? // Вопросы прикладной лингвистики. – Выпуск № 7. – М.: Изд-во РУДН, 2012. – С. 48-54.

56.Назарова, Т.Б. Словарь общеупотребительной терминологии английского языка делового общения. – М.: Астрель/АСТ, 2002.

57.Николаева, Т.М. Лингвистика текста // Лингвистический энциклопедический словарь / гл. ред. В.Н. Ярцева. – М., 1990.

58.Николаева, Т.М. Лингвистика текста: Современное состояние и перспективы // Новое в зарубежной лингвистике. М., 1978. – Вып. 8.

59.Папина, Ф.А. Текст: его единицы и глобальные категории. – М.: Едиториал УРСС, 2002.

60.Плунгян, В.А. Почему современная лингвистика должна быть лингвистикой корпусов? (Публичная лекция, прочитанная 01.10.2009). [Электронный ресурс]. – Режим доступа: http://www.polit.ru/article/2009/10/23/corpus/. Проверено 01.10.18.

61.Разинкина, Н.М. Функциональная стилистика. – М.: Высшая школа, 2004.

62.Ревзин, И.И. Современная структурная лингвистика. – М.: Наука, 1977.

63.Рыков, В.В. Курс лекций по корпусной лингвистике. [Электронный ресурс]. – Режим доступа: http://rykov-cl.narod.ru/c.html. Проверено 01.10.18.

64.Рымарь, Н.Т. Теория автора и проблема художественной деятельности / Н.Т. Рымарь, В.П. Скобелев – Воронеж, 1994.

65.Савицкий, В.М. Основы общей теории идиоматики. – М.: Гнозис, 2006.

66.Салимовский, В.А. Жанры речи в функционально-стилистическом освещении (научный академический текст). – Пермь: Издательство Пермского университета, 2002.

67.Седов, К.Ф. Жанр и коммуникативная компетенция // Хорошая речь. – Саратов: Издательство Саратовского университета, 2001.

68.Степанов, Ю.С. Константы. Словарь русской культуры. Опыт исследования. – М.: Языки русской культуры, 1997.

69.Стилистический энциклопедический словарь русского языка / под ред. М.Н. Кожиной. – М.: Флинта: Наука, 2003.

70.Теория литературных жанров / под ред. Н.Д.Тамарченко. – М.: Академия, 2012.

71.Толстова, Т.В. Электронный корпус англоязычной деловой корреспонденции VOBEC // Материалы конференции «Интеграция международных процессов в образовании, бизнесе и межкультурной коммуникации», 20-21 марта 2007 г. – Самара, ПФ МУМ, 2007. – С. 102-113.

72.Тураева, З.Я. Лингвистика текста (текст: структура и семантика). – М.: Просвещение, 1986.

73.Федосюк М.Ю. Нерешенные вопросы теории речевых жанров // Вопросы языкознания. 1997. № 5. – С. 102–120.

74.Чернявская, В.Е. Дискурс как объект лингвистических исследований // Текст и дискурс. Проблемы экономического дискурса: Сб. науч. тр. – СПб.: Изд-во С.-Петерб. гос. ун-та экономики и финансов, 2001. – С. 14–17.

75.Шкловский, В.Б. Кончился ли роман? // Иностранная литература. 1967, № 8. – С. 218-231.

76.Шмелева, Т.В. Жанроведение? Генристика? Генология? // Антология речевых жанров: повседневная коммуникация / под ред. проф. К.Ф. Седова. – М.: Лабиринт, 2007. – С. 62-67.

77.Шмелева, Т.В. Несложившаяся традиция отечественной филологии // Филология – Журналистика ’94: Научные материалы. – Красноярск, 1995.

78.Шмелева, Т.В. Речевой жанр. Возможности описания и использования в преподавании языка // Russistik. Berlin. 1992. № 2. – С. 20-32.

79.Щукин, А.Н. Обучение иностранным языкам: Теория и практика. – М.: Филоматис, 2004.

80.Aarts, J. Intuition-Based and Observation-Based Grammars // English Corpus Linguistics / Ed. by K. Aijmer, B. Altenberg. – London: Longman, 1991. – P. 44-62.

81.Anthony, L. A critical look at software tools in corpus linguistics // Linguistic Research. № 30(2), 2013. – P. 141–161.

82.Aston, G. and Burnard, L. The BNC Handbook. Edinburgh, U.K.: Edinburgh University Press, 1998.

83.Atkins, S., Clear, J., and Ostler, N. Corpus design criteria // Literary and Linguistic Computing. № 7(1), 1992. P. 1–16.

84.Baker, P., Hardie, A., McEnery, T. A Glossary of Corpus Linguistics. – Edinburgh: Edinburgh University Press, 2006.

85.Bargiela-Chiappini, F. Meaning Creation and Genre across Cultures: Human Resource Management Magazines in Britain and Italy // Writing Business: Genres, Media, and Discourses / Ed. by F. Bargiela-Chiappini and C. Nickerson. – Pearson Education Limited, 1999. – P. 129-152.

86.Baroni, M. Distributions in text // Corpus Linguistics: An International Handbook / Ed. by A. Ludeling, M. Kyto. – Vol. 2. – Berlin, Germany: Mouton de Gruyter, 2009. – P. 803–822.

87.Bennet, G.R. Using Corpora in the Language Learning Classroom: Corpus Lingsuistics for Teachers. – Ann Arbor, Michigan: University of Michigan Press, 2010.

88.Bhatia, V.K. Analysing Genre: Language Use in Professional Settings (Applied Linguistics and Language Study). – Longman, 1993.

89.Bhatia, V.K. Analysing Genre: Some Conceptual Issues // Academic Writing in Context: Implications and Applications / Ed. By M.Hewings. – University of Birmingham Press, 2001.

90.Bianchi, F. Culture, corpora and semantics: Methodological issues in using elicited and corpus data for cultural comparison. – Università del Salento, 2012.

91.Biber D., Conrad, S., Reppen, R. Corpus Linguistics: investigating language structure and use. – Cambridge: Cambridge University Press, 1998.

92.Biber, D. Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation // Literary and Linguistic Computing. № 5, 1990. – P. 257-269.

93.Biber, D. Representativeness in corpus design // Literary and Linguistic Computing. № 8(4), 1993. – P. 243–257.

94.Bowker, L., Pearson, J. Working with specialized language. London, New York: Routledge, 2002.

95.Braun, S. From Pedagogically Relevant Corpora to Authentic Language Learning Contents // ReCALL. № 17(1), 2005. P. 47–64.

96.Brazil, D. A Grammar of Speech. – Oxford University Press, 1995.

97.Brezina, V. Statistics in Corpus Linguistics: A Practical Guide. – Cambridge University Press, 2018.

98.Buendgens-Kosten. J. Authenticity // *ELT Journal.* – *№* 68(4). – P. 457–459.

99.Carter, R., McCarthy, M. Cambridge Grammar of English. A Comprehensive Guide to Spoken and Written English Grammar and Usage. – Cambridge University Press, 2006.

100. Chomsky, N. Aspects of the theory of syntax. – Cambridge, Massachussets: MIT Press, 1965.

101. Corpus Lingstics 2013. Abstract Book / Ed. by A.Hardie, R. Love. – Lancaster: UCREL, 2013.

102. Corpus Linguistics around the World: / Ed. by A.Wilson, D. Archer, P.Rayson. – Amsterdam-New York: Rodopi, 2016.

103. Corpus Linguistics: An International Handbook / Ed. By A.Lüdeling, M.Kytö. –·Berlin-New York: Walter de Gruyter, 2008.

104. Crystal, D. The Cambridge Encyclopedia of the English Language. – Cambridge University Press, 1995.

105. Developing linguistic corpora: a guide to good practice / Ed. by M. Wynne. – Oxford: Oxbow Books, 2005. [Электронный ресурс]. – Режим доступа: http://ahds.ac.uk/linguistic-corpora/. Проверено 01.10.18.

106. Dudley-Evans, T., St John, M.J. Developments in ESP. A Multi-Disciplinary Approach. – Cambridge University Press, 2003. – P. 75 – 76.

107. Emerson, P. Email English. – Macmillan, 2005.

108. Evans, D. Powerhouse: An Intermediate Business English Course. – Oxford: Longman, 1999.

109. Firth, J.R. Papers in Linguistics 1934–1951. – London: Oxford University Press, 1957.

110. Francis, W.N., Kučera, H. Frequency Analysis of English Usage: Lexicon and Grammar. – Boston: Houghton Mifflin. 1982.

111. Garside, R. and Leech, G. Running a grammar factory: the production of syntactically analysed corpora or “treebanks”’ // English Computer Corpora: Selected Papers and Research Guide / Ed. by S.

Johansson and A.-B. Stenström – Berlin-New York: Mouton de Gruyter, 1991. – P.15-32.

112. Greenbaum, S., Quirk, R. A Student's Grammar of the English Language. – Longman, 1990.

113. Gries, S. Th. Corpus-linguistics and theoretical linguistics: A love-hate relationship? Not Necessarily". – International Journal of Corpus Linguistics № 15(3). – P. 327–343

114. Gries, S.Th. What is Corpus Linguistics? // Language and Linguistics Compass. – № 3, 2009. – P. 1–17.

115. Habermas, J. Theorie des kommunikativen Handelns. – Frankfurt am Main: Suhrkamp, 1981.

116. Halliday, M.A.K. and Matthiessen C.M.I.M. An Introduction to Functional Grammar. – London: Arnold, 2004.

117. Halliday, M.A.K., Hasan, R. Cohesion in English. – Longman, 2001.

118. Hofstede, G. Culture's Consequences, Comparing Values, Behaviors, Institutions, and Organizations across Nations. – CA: Sage Publications, 2001.

119. Hunston, S. Corpora in Applied Linguistics. – Cambridge: Cambridge University Press, 2002.

120. Hunston, S. Corpus linguistics // Encyclopedia of language and linguistics / Ed. by K. Brown. – Amsterdam: Elsevier, 2006. – P. 234–248.

121. Hutchinson, T., Waters, A. English for Specific Purposes: A Learner-centred Approach. – Cambridge University Press, 1995.

122. Johns, A.M. Text, Role and Context: Developing Academic Literacies. – Cambridge: Cambridge University Press, 1997.

123. Johns, T. Data-driven learning: The perpetual challenge// Teaching and learning by doing corpus linguistics / Ed. by B. Kettermann, G. Marko. – Amsterdam: Rodopi, 2002. – P. 107–117.

124. Jones, L., Alexander, R. New International Business English: Teacher's Book. – Cambridge University Press, 2001.

125. Karlsson, F., Voutilainen, A., Heikkilä, J., Anttila, A. Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text. – Berlin, Germany: Mouton de Gruyter, 1995.

126. Kennedy, G. An Introduction to Corpus Linguistics. London-New York, 1998.

127. Kirkpatrick, A. Information Sequencing in Mandarin Letters of Request // Anthropological Linguistics. – No 33. 1991. – P. 183-203.

128. Knowles, G. Corpora, databases and the organization of linguistic data // Using corpora for language research / Ed. by J. Thomas, M. Short. – London: Longman, 1996. – P. 14-26.

129. Lee, D. Y. W. Genres, registers, text types, domains and styles: Clarifying the concepts and nevigating a path through the BNC jungle // Language Learning and Technology. – № 5(5), 2001. – P. 37–72.

130. Lee, D. Y. W. Genres, Registers, Text Types, Domains, and Styles // Language Learning and Technology. – № 5(3), 1997. – P. 37–72.

131. Leech, G. Introducing Corpus Annotation // Corpus Annotation / Ed. by R. Garside, G. Leech, and A. McEnery. – London, U.K.: Longman, 1997. – P. 1–18.

132. Leech, G. The state of the art in corpus linguistics // English Corpus Linguistics: Linguistic Studies in Honour of Jan Svartvik. – London: Longman, 1991. – P. 8-29.

133. Leech, G., Svartvik J. A Communicative Grammar of English. – Longman, 1994.

134. Locker, K.O. Business and Administrative Communication. – Irwin McGraw-Hill, 1998.

135. Longman Dictionary of Contemporary English / Ed. by Randolf Quirk, Della Summers. – Longman, 1978.

136. Longman Grammar of Spoken and Written English / Ed. by Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan. – Pearson Education Limited, 1999.

137. Longman Language Activator: Helps You Write and Speak Natural English / Ed/ by Addison Wesley Longman, Trudy Longman. – Pearson-Longman, 2000.

138. Macmillan English Dictionary for Advanced Learners. – Macmillan, 2002.

139. Mascull, B. Key Words in Business. – Collins COBUILD, 1996.

140. Mascull, B. Key Words in Science and Technology. – Collins COBUILD, 1997.

141. Mascull, B. Key Words in the Media. – Collins COBUILD, 1995.

142. McArthur, T. Longman Lexicon of Contemporary English. – London: Longman, 1981.

143. McEnery, A. Corpus linguistics // The Oxford Handbook of Computational Linguistics / Ed. by in R. Mitkov. – Oxford: Oxford University Press, 2003. – P. 448-463.

144. McEnery, A. M., Wilson, A. Corpus Linguistics: An Introduction. – Edinburgh: Edinburgh University Press, 2001.

145. McEnery, A.M., Xiao, R.Z, Tono, Y. Corpus-based language studies: An advanced resource book. Routledge Applied Linguistics Series. London: Routledge, 2006.

146. McEnery, T., Hardie, A. Corpus linguistics: Method, theory and practice. – Cambridge: Cambridge University Press, 2012.

147. Meyer, C.F. English Corpus Linguistics: An Introduction. – Cambridge: Cambridge University Press, 2002.

148. Nazarova, T.B. Business English. A Course of Lectures and Practical Assignments. – M.: AST/Astrel, 2004.

149. Nelson, M. A Corpus-Based Study of Business English and Business English Teaching Materials. PhD Thesis. – Manchester: University of Manchester, 2000. – . [Электронный ресурс]. – Режим доступа: http://users.utu.fi/micnel/business_english_lexis_site.htm . Проверено 01.10.18.

150. Nelson, M. Building a written corpus: what are the basics? // The Routledge handbook of corpus linguistics / Ed. by A. O'Keeffe, M. Michael McCarthy. – Routledge, 2010. – P. 53–65.

151. O'Keeffe, A., McCarthy, M., Carter, R. From Corpus to Classroom. – Cambridge: Cambridge University Press, 2007.

152. Orlikowski, W., Yates, J. Genre Repertoire: The Structuring of Communicative Practices in Organizations // Administrative Science Quarterly 39. 1994. – P. 541-574.

153. Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. A Comprehensive Grammar of the English Language. – Longman, 1985.

154. Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. A Grammar of Contemporary English. – Longman, 1972.

155. Rayson, P., Archer, D., Piao, S., McEnery, A. M. *The UCREL semantic analysis system //* Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop. – Lisbon, Portugal: 2004. – P. 7-12.

156. Rippen, R. Building a corpus: what are the key considerations? // The Routledge handbook of corpus linguistics / Ed. by A. O'Keeffe, M. Michael McCarthy. – Routledge, 2010. – P. 31–37.

157. Schiffrin, D. Approaches to Discourse. – Blackwell, 1994.

158. Scott, M. WordSmith Tools. – Oxford: Oxford University Press, 1999.

159. Scott, M., Tribble, C. Textual Patterns: Key Words and Corpus Analysis in Language Education. – Philadelphia: John Benjamins, 2006.

160. Simpson, R., Swales, J.M. Corpus Linguistics in North America. – Ann Arbor: The University of Michigan Press, 2001.

161. Sinclair, J. Collins COBUILD English Grammar. – HarperCollins, 1990.

162. Sinclair, J. Corpus and text – basic principles // Developing linguistic corpora: a guide to good practice / Ed. by M. Wynne. – Oxford: Oxbow Books, 2005. – P. 1–16. [Электронный ресурс]. – Режим доступа: http://ahds.ac.uk/linguistic-corpora/. Проверено 01.10.18.

163. Sinclair, J. Corpus, Concordance, Collocation. – Oxford: Oxford University Press, 1991.

164. Sinclair, J. EAGLES. Preliminary Recommendations on Corpus Typology, 1996. [Электронный ресурс]. – Режим доступа: http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html. Проверено 01.10.18.

165. Stubbs, M. Discourse Analysis: The Sociolinguistic Analysis of Natural Language. – Blackwell Publisher, 1989.

166. Swales J.M. Research Genres: Exploration and Applications. – Cambridge University Press, 2004.

167. Swales, J.M. Aspects of Article Introductions. – Aston ESP Research Reports No. 1. – Birmingham: The University of Aston, 1981.

168. Swales, J.M. Genre Analysis: English in Academic and Research Settings. – Cambridge University Press, 1990.

169. Tognini-Bonelli, E. Corpus Linguistics at Work. – Amsterdam: John Benjamins Publishing Company (Studies in corpus linguistics), 2001.

170. Upton, T. A., Connor, U. Using Computerized Corpus Analysis to Investigate the Textlinguistic Discourse Moves of a Genre // English for Specific Purposes. – № 20, 2001. – P. 313–329.

171. Using Corpora to Explore Linguistic Variation / Ed. by R.Reppen, S.M.Fitzmaurice, D.Biber. – Amsterdam/Philadelphia: John Benjamins Publishing Company, 2002.

172. Widdowson, H.G. Explorations in Applied Linguistics. – Oxford: Oxford University Press, 1979.

173. Wilson, A., Thomas, J. Semantic Annotation // Corpus Annotation / Ed. by R. Garside, G. Leech, and A. McEnery. – London, U.K.: Longman, 1997. – P. 53-65.

174. Xiao, R. Well-known and influential corpora: A survey // Corpus Linguistics: An International Handbook. – Berlin: Mouton de Gruyter, 2008. – Vol. 1. – P. 383–457.

Научное издание

Толстова Татьяна Витальевна

ЖАНР И КОРПУС: СОВРЕМЕННЫЕ ПОДХОДЫ К ИЗУЧЕНИЮ И ПРЕПОДАВАНИЮ ЯЗЫКА

Монография

Редактор М.С. Сараева
Компьютерная верстка И.И. Спиридоновой

Подписано в печать 04.12.2018. Формат 60 × 84 1/16.
Бумага офсетная. Печ. л. 12,25.
Тираж 300 экз. (1 з-д 1-30). Заказ .

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САМАРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА»
(САМАРСКИЙ УНИВЕРСИТЕТ)
443086, САМАРА, МОСКОВСКОЕ ШОССЕ, 34.

Изд-во Самарского университета.
443086, Самара, Московское шоссе, 34.