УДК 81.33

ПРОЛЕГОМЕНЫ К КОРПУСНОЙ ЛИНГВИСТИКЕ

Захаров Виктор Павлович
Доцент кафедры математической лингвистики
Санкт-Петербургский государственный университет
С-Петербург, Университетская наб., д. 7/9.
e-mail: v.zakharov@spbu.ru

Статья содержит описание предмета и основного содержания корпусной лингвистики. Представлен небольшой исторический экскурс. Дается понятие корпуса и характеризуются основные понятия корпусной лингвистики: репрезентативность, сбалансированность, разметка, корпусный менеджер и др. Представлены основания для типологии и основные типы корпусов. Вопрос о предмете корпусной лингвистики, ее месте среди других разделов и направлений в языкознании встал и обострялся по мере того, как развивались корпусные технологии. Сейчас корпусная лингвистика все время расширяет поле соприкосновения с теоретической лингвистикой и с разными направлениями внутри нее - поэтому важно осознать ее место в в рамках лингвистической науки. Эта тема также затронута в работе. Обозначены перспективы развития, как внутри направления, так и совместно с другими лингвистическими дисциплинами.

Ключевые слова: корпусная лингвистика, корпус, разметка, репрезентативность, прагматика, лингвистическая теория, статистические данные

Посвящается Наталье Владимировне Уфимцевой,
моей однокурснице по учебе в Ленинградском университете

Введение

Что такое корпусная лингвистика? Этот простой и одновременно непростой вопрос стал занимать лингвистов, когда корпусы вошли в широкий лингвистический оборот. Раздел лингвистики? Тогда какой аспект языка она изучает? Раздел компьютерной лингвистики? Методология лингвистического исследования? Лингвистическая теория? Однозначного ответа на этот вопрос нет до сих пор, но его постановка помогает лучше понять предмет данного феномена. Корпусная лингвистика не является монолитным набором категорий или методов, это некое гетерогенное поле, внутри которого варьируются различные подходы к созданию и использованию корпусов.

Первые лингвистические корпусы текстов появились в 60-е гг. прошлого столетия. В 1963 г. в Брауновском университете (США) впервые был создан большой корпус текстов на машинном носителе (Brown Corpus). Авторы корпуса, У. Френсис (W. Francis) и Г. Кучера (H. Kučera), спроектировали его как набор из пятисот 2000-словных прозаических печатных текстов американского варианта английского языка. Тексты принадлежали пятнадцати наиболее массовым жанрам англоязычной печатной прозы США и были напечатаны в 1961 г. Корпус сопровождался большим количеством материалов его первичной статистической обработки – частотный и

алфавитно-частотный словарь, разнообразные статистические распределения. Появление Брауновского корпуса вызвало всеобщий интерес и оживленные дискуссии. Прежде всего они коснулись принципов отбора текстов и состава потенциально решаемых на таком корпусе задач. Затем последовал Ланкастерский корпус английского языка (Lancaster-Oslo-Bergen Corpus, LOB), представляющий собой своего рода аналог Брауновского корпуса для английского языка, и много других корпусов для разных языков. В первой половине 1990-х гг. корпусная лингвистика окончательно сформировалась как новое лингвистическое направление.

Но правильно ли будет сказать *новое*?

Ключевая и знаменитая фраза мольеровского героя из «Мещанина во дворянстве» звучит так: «…я и не подозревал, что вот уже более сорока лет говорю прозой». Открытие, сделанное господином Журденом, должно изобличать, конечно, его безграмотность, однако можно сказать, что мы действительно говорим прозой. Так же можно сказать, что лингвисты всю жизнь занимались корпусной лингвистикой, не подозревая об этом. Не случайно одна из статей У. Фрэнсиса называется Corpora B.C. (корпусы до Р.Х.) [Francis 1991]. Здесь, конечно, игра слов: автор имел в виду *корпусы до компьютеров (*Corpora Before Computers*)*. Но в статье речь идет о том, что идеи корпусной лингвистики действительно зародились в докомпьютерную эпоху. Лингвисты и лексикографы уже давно в своей работе используют эмпирический материал, цитаты из текстов, которые выписывались на карточки и образовывали «корпусы» под названием картотеки.

Основной выходной продукт из корпусов – это конкорданс, но и он «изобретен» давным-давно. Первая «конкорданция» появилась в начале XIII века («Concordantiae morales sacrae scripturae», «Нравственная конкорданция Священного Писания»). Это был своего рода предметный указатель к текстам Библии. Следом около 1230 года появилась конкорданция к Вульгате Гуго де Сен-Шера, первого кардинала доминиканского монастыря святого Иакова в Париже (Concordantiae Sancti Jacobi). Для составления ее Сен-Шер воспользовался услугами 500 доминиканцев, собратий своего монастыря. Конкорданция при каждом слове приводила цитаты из Библии, с указанием места, где они находятся.

И все же будет правильным сказать, что корпусная лингвистика – это наука XXI века. Сегодня это неотъемлемая часть лингвистики, можно сказать, ее «тело», а «двигателем» является компьютерная лингвистика [Плунгян 2008].

Можно сказать, что все современные лингвистические исследования и работы по составлению словарей и грамматик так или иначе ориентированы на использование представительных корпусов текстов. Развитие современных интеллектуальных программных систем, предназначенных для обработки текстов на естественном языке, также требует большой экспериментальной лингвистической базы. Спрос на корпусные данные совпал с появлением соответствующих технических возможностей.

Помимо возможности получения обширного текстового материала, релевантного запросу (конкорданс), корпусы могут использоваться для получения разнообразных справок и статистических данных о языковых и речевых единицах. В частности, на основе корпусов можно получить данные о частоте словоформ, лексем, грамматических категорий, проследить изменение частот и контекстов в различные периоды времени, получить данные о совместной встречаемости лексических еди-

ниц и т.д. Репрезентативный массив языковых данных за определенный период позволяет изучать динамику процессов изменения лексического состава языка, проводить анализ лексико-грамматических характеристик в разных жанрах и у разных авторов, и т.д.

Конкорданс и статистические данные представляют, соответственно, две формы анализа, качественную и количественную, и обе являются «краеугольными камнями» корпусной лингвистики.

1. Основные понятия корпусной лингвистики

Основное понятие корпусной лингвистики, как нетрудно догадаться – это **корпус**. Существует множество определений понятия «корпус». Но, наверное, во всех определениях будет указано на 4 главные его «ипостаси»: 1) электронный; 2) репрезентативный; 3) размеченный; 4) прагматически ориентированный (создаваемый для определенных целей). Поясним вышеназванные «ипостаси».

Электронный

То, что корпус в современном понимании всегда электронный, сразу ставит вопрос о специальной программной системе для работы с ним. Корпус текстов становится мощным инструментом в руках лингвиста лишь ее посредством. Эту систему называют *корпусным менеджером* (или *корпус-менеджером*) (англ. corpus manager). Это специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления пользователю результатов в удобной форме.

Поиск в корпусе позволяет по любому слову построить *конкорданс* (concordance) – список всех употреблений данного слова в контексте со ссылками на источник. Поэтому раньше эту поисковую программу называли *конкордансер*.

Современный корпусный менеджер представляет собой мощную программно-лингвистическую систему, позволяющую, помимо выдачи конкорданса и простой статистики, решать сложные задачи, такие как построение разнообразных частотных списков, выявление коллокаций (устойчивых словосочетаний с указанием силы связи между компонентами), выявление коллокаций с учетом синтаксических формул, выявление ключевых слов и словосочетаний, построение лексико-семантических групп (тезаурусов) и др.

Репрезентативный

Термин *репрезентативность* (*англ*. representativeness) можно перевести как представительность, т.е. корпус должен хорошо представлять тот объект, который он моделирует, в общем случае, весь язык [Biber 1993]. Но могут создаваться и корпусы, «представляющие» какой-то подъязык, например, корпус русского языка первой половины XX века, или корпус русского сонета, или корпус разговорной речи жителей Заполярья. Почти в любом случае проблема репрезентативности существует и ее надо решать. Есть и исключения. Допустим, при создании корпуса писем Н.В. Гоголя очевидно, что туда должны войти все письма Н.В. Гоголя. Но это, скорее, исключение. Естественно и понятно, что корпус – это собрание текстов конечного фиксированного размера. В реальном языке или подъязыке таких текстов, как правило, гораздо больше. Тогда встают две проблемы: проблема объема и проблема отбора.

Проблема объема явно была сформулирована в 1960-70-е гг. XX в. при создании частотных словарей, когда обсуждалось понятие представительной выборки

– такого количества языкового материала, после достижения которого относительные частоты языковых единиц практически не меняются. Объем первых корпусов составлял 1 млн словоупотреблений (Брауновский корпус, корпус Ланкастер-Осло-Берген, корпус Частотного словаря русского языка под ред. Л.Н. Засориной). Такой объем не позволял отразить язык во всем его многообразии. Затем стали считать, что общеязыковой (национальный) корпус должен включать не менее 100 млн словоупотреблений. Очевидно, что для изучения многих языковых явлений и этого объема недостаточно. Поэтому сейчас создаются корпусы, где счет словоупотреблениям идет на миллиарды.

Вторая проблема формирования представительной выборки - проблема отбора. Из каких текстов сформировать тот самый минимально необходимый объем? Поэтому появилось еще одно важное понятие корпусной лингвистики – *сбалансированность корпуса* (англ. balance). Эту характеристику обеспечить еще труднее, особенно применительно к национальным корпусам. Если корпус – это уменьшенная модель языка, то значит, в нем пропорционально должны быть представлены текстов различных периодов, жанров, стилей, авторов и т. д. Можно сказать, что применительно к языку в целом мы этих пропорций не знаем. Тем не менее, к этому нужно стремиться, как на этапе проектирования корпуса, так и на этапе его развития.

Именно репрезентативность и сбалансированность корпуса определяют достоверность полученных на его материале результатов.

Размеченный

Разметка корпуса, или аннотация (tagging, annotation) – это главное, что отличает лингвистический корпус от коллекции текстов или электронной библиотеки. Разметка заключается в приписывании текстам корпуса и их компонентам дополнительной информации, метаданных [Atkins S. et al. 1992]. Метаданные можно поделить на 3 типа: экстралингвистические, данные о структуре текста, лингвистические. При выборе этих данных необходимо руководствоваться целями исследования и потребностями лингвистов, а также возможностями по внесению в текст тех или иных дополнительных признаков. В целом рекомендуется использовать следующие основополагающие принципы:

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

Эти принципы перекликаются с 7 постулатами аннотирования, сформулированными в 1993 г. английским лингвистом Дж. Личем (Leech's seven maxims of annotation) [Leech 1993].

Экстралингвистическая разметка (метаразметка) включает в себя «внешнюю», «интеллектуальную» разметку текстов корпуса (библиографические характеристики, типологические, тематические, социологические характеристики). Например, художественные тексты в Национальном корпусе русского языка (НКРЯ) подразделяются по жанрам: детектив, боевик, детская, документальная проза, драматургия, историческая проза, любовная история, нежанровая проза, приключения, фантастика, юмор и сатира. Нехудожественная литература делится по сферам

функционирования: бытовая, официально-деловая, производственно-техническая, публицистика, реклама, учебно-научная, церковно-богословская, электронная коммуникация. Набор признаков для метаданных чаще всего основывается на рекомендациях проекта TEI (Text Encoding Initiative). Экстралингвистическая разметка нужна, во-первых, для выявления взаимосвязи языка и условий его существования; во-вторых, для отбора и изучения отдельных подмножеств языка.

Среди лингвистических типов разметки выделяются:

- *морфологическая* разметка. В иностранной терминологии употребляется термин part-of-speech tagging (POS-tagging), дословно – частеречная разметка. В действительности морфологические метки включают не только признак части речи, но и признаки грамматических категорий, свойственных данной части речи. На сегодняшний день это основной тип разметки: во-первых, большинство крупных корпусов являются как раз морфологически размеченными корпусами, во-вторых, морфологический анализ рассматривается как основа для дальнейших форм анализа – синтаксического и семантического, и, в-третьих, успехи в автоматическом морфологическом анализе позволяют размечать корпусы больших размеров;

- *синтаксическая* разметка, являющаяся результатом синтаксического анализа, или *парсинга* (англ. parsing), выполняемого на основе данных морфологического анализа. Этот вид разметки описывает синтаксические связи между лексическими единицами и различные синтаксические конструкции (например, придаточное предложение, глагольное словосочетание и т.п.);

- *семантическая* разметка. Хотя для семантики нет единой семантической теории, чаще всего семантические тэги обозначают семантические категории, к которым относится данное слово или словосочетание, и более узкие подкатегории, специфицирующие его значение;

Существует много и других типов разметки.

Прагматически ориентированный

Практика разработки и применения электронных корпусов текстов показала, что невозможно создать универсальный корпус. Задачи и цели любого исследования определяют тип корпуса, правила отбора текстов и способ и степень их обработки. Корпусы всегда создаются под определенную задачу или круг задач. Эта задача определяет как наполнение корпуса текстами (например, письма Н.В. Гоголя, русская драма 19-го века, тексты языка охотников), так и разметку корпуса (морфологическая, синтаксическая и т.д., с учетом разнообразия жанров, с учетом социологических характеристик говорящих или без). Корпусная прагматика включает в себя также и аудиторию, для которой создается корпус.

Постепенно сформировалась терминология корпусной лингвистики [Baker et al. 2006] и в настоящее время идет формирование соответствующнй русскоязычной терминологии.

2. Предмет корпусной лингвистики

Задаваясь вопросом о месте корпусной лингвистике в лингвистике вообще, видимо, правильнее всего будет сказать, что это методология лингвистического исследования, применимая практически к любой области лингвистики. Однако существует и другой взгляд, заключающийся в том, что корпусная лингвистика должна развиваться как отдельное направление, черпающее теорию преимущественно из

корпуса. В англоязычной литературе эти подходы получили название *corpus-based* (корпусно-ориентированный) и *corpus-driven* (что можно перевести как корпусно-управляемый).

Первый подход предполагает, что корпусы используются для проверки лингвистических теорий или гипотез, чтобы их подтвердить, опровергнуть или уточнить. Этот подход достаточно традиционен [Xiao, Tono 2006]. Второй подход провозглашает, что корпус сам является главным и чуть ли не единственным источником наших знаний о языке. Здесь корпусная лингвистика получает статус теории («Теория не существует независимо от данных» Tognino-Bonelli 2001: 84-85]) и рассматривается как «важнейший концепт в лингвистической теории» [Stubbs 1993: 24]. Утверждается, что корпус неявно содержит в себе теорию языка и нужно ее оттуда только «добыть». Такое понимание лингвистики как чисто эмпирической науки возвращает нас к работам американских структуралистов первой трети XX в.

В недрах корпусной лингвистике этот подход называют нео-фёрсианским (neo-Firthian), связывая его с именем Дж.Р. Фёрса (J.R. Firth). Фёрс ввел в научный обиход понятие коллокации. Может быть, самой знаменитой цитатой в корпусной лингвистике является его высказывание «Вы поймете слово по его окружению» (“You shall know a word by the company it keeps”) [Firth 1957: 179]. Утверждается, что значение слова (равно как и другие лингвистические концепты) существует только в потенции и что реально в языке оно возникает только в контексте. Эти взгляды Фёрса приобрели популярность и получили дальнейшее развитие в работах корпусных лингвистов 1990-2000 гг. Предполагается, что аналитик, исследующий данные, не использует никаких априори установленных теоретических концепций [Sinclair 1991, 2004]. На самом деле это трудно себе представить, но такая «индуктивная» точка зрения постулируется.

Другой краеугольный камень подхода нео-фёрсианцев к изучению языка – это понятие дискурса. Дискурс для них – это не только текст, «практика» языка, но и способ реализации самого языка или подъязыка, не только способ говорения, но и способ мышления. И здесь воззрения ряда ученых, исповедующих это направление и использующих корпусные ресурсы и методы, в ряде позиций стыкуются с психолингвистикой и с социолингвистикой [Sinclair 2004]. Термин «корпусно-управляемый подход» также часто используется для обозначения любых индуктивных исследований, использующих так называемые «сырые» (raw) корпусы без лингвистической разметки.

На наш взгляд, эта концепция корпусно-управляемых исследований слишком умозрительна и часто расходится с практикой самих лингвистов, ее исповедующих. Теория всегда рождается как акт человеческого взаимодействия с действительностью (с данными), но при этом познающий субъект обладает сознанием, предрассудками, включенностью в определенные социальные системы, включая науку.

Поэтому де факто корпусная лингвистика, как она есть, всегда корпусно-ориентированная. Другое дело, что широкое использование корпусных ресурсов и методов в современной лингвистике стимулирует лингвистов к тому, чтобы придать данным повышенный вес, и создает для этого соответствующие возможности.

Итак, можно сказать, что корпусная лингвистика имеет своим предметом теоретические основы и практические механизмы, обусловленные математическими

методами и компьютерными технологиями, создания и использования представительных массивов языковых данных, предназначенных для лингвистических исследований в интересах широкого круга пользователей.

3. Типы корпусов

Несмотря на разнообразие корпусов, можно выделить два основных способа деления корпусов на классы. Первый способ – это противопоставление корпусов, относящихся ко всему языку (часто к языку определенного периода), корпусам, относящимся к какому-либо подъязыку (жанр, стиль, язык определенной возрастной или социальной группы, язык писателя или ученого и т.п.). Второй способ – разделение корпусов по типу лингвистической разметки. Несмотря на наличие множества типов разметки, большинство реально существующих корпусов относится к корпусам морфологического либо синтаксического типа (последние в англоязычной литературе называют treebanks, что можно перевести как «банки синтаксических структур»). При этом следует подчеркнуть, что корпус с синтаксической разметкой явно или неявно включает в себя и морфологические характеристики лексических единиц.

С точки зрения типа корпусных данных главное разделение следует провести между корпусами письменного языка (их подавляющее большинство) и корпусами устной речи. Создание репрезентативного корпуса устной речи является сложной и трудоемкой задачей. Построение корпусов устной речи продвигается намного медленнее, чем построение корпусов письменной речи. В первую очередь устную речь нужно зафиксировать на каком-то носителе. Далее встает трудоемкая задача ее транскрибирования, а также маркирования в составе фонетического (устного, речевого) корпуса паралингвистических явлений, сопутствующих речи (паузы, смех, бормотание, кашель и т. п.). Несмотря на трудности создания, в мире создается достаточно много речевых корпусов. В качестве примеров таких корпусов можно привести корпус «Один речевой день» (ОРД), разрабатываемый в Санкт-Петербургском университете [Шерстинова и др. 2009], корпус устной речи НКРЯ [Гришина, Савчук 2009], мультимедийный корпус МУРСО в составе НКРЯ, включающий, кроме фонетики, еще и видеоряд и аннотацию жестов [Гришина 2009].

С точки зрения режима создания и пополнения корпуса выделяют статические корпусы (static, balanced, sample corpora) и мониторные корпусы (monitor corpora). Первые – это фиксированные массивы, представляющие срез языка определенного периода, «сбалансированный» в соответствии с принятыми критериями (Brown corpus, LOB, British National Corpus). Вторые – это регулярно пополняемые корпусы (НКРЯ, Corpus of Contemporary American, Bank of England), при этом «баланс» корпуса может меняться.

Вообще же существует большое число различных типов корпусов. Их разнообразие определяется многообразием исследовательских и прикладных задач, для решения которых они создаются, и различными основаниями для классификации.

4. Web as corpus

Но хотя мониторные корпусы и достигают больших размеров, существует еще бóльший, постоянно растущий корпус – это World Wide Web. Поэтому появились идеи использовать поисковые массивы и системы Интернета для решения лингвистических задач [Беликов 2004; Захаров 2005]. Были также проекты создания специальных поисковых машин-посредников, имеющих корпусный интерфейс, но пользующихся базами данных поисковых систем (см., например, WebCorp [Renouf 2003]). Однако в целом этот путь оказался малопродуктивным и периферийным.

Затем возникла идея создавать полноценные корпусы на основе текстов, взятых из Интернета. Вероятно, первым ее высказал английский лексикограф Адам Килгаррифф (A. Kilgarriff) в 2001 г. [Kilgarriff 2001; Kilgarriff, Grefenstette 2003]. В этом случае создание очень больших корпусов уже не кажется проблемой. Источники текстов разнообразны и бесконечны. Но какие тексты брать? Принципиальной разницей между поиском в вебе и извлечением веб-данных для корпуса является то, что в первом случае мы знаем, что ищем, в то время как во втором мы пытаемся найти что-то не очень четко определенное [Беликов 2013]. Вначале технология создания корпусов на основе текстов из веба столкнулась с большими трудностями как технического, так и идеологического характера [Sharoff 2006]. Однако многие из них уже решены и за прошедшее десятилетие эта технология, получившая название *WaCky*, достигла заметных успехов [Baroni et. al 2009; Benko 2014; Jakubíček et. al. 2013]. При этом в автоматическом режиме приходится решать задачи, связанные как с особенностями текстов из веба (обилие ошибок, дублирование информации и т.п.), так и собственно корпусные (сбалансированность, разметка). Веб-корпусы не заменяют и не отменяют традиционные, но они их существенно дополняют, и уже сегодня создаются корпусы объемом несколько десятков миллиардов словоупотреблений, позволяющие изучать широкую периферию языка и изменения в нем.

Заключение

Итак, корпусная лингвистика представляет собой новое направление в лингвистической науке, позволяющее проводить исследование единиц любого языкового уровня в реальном их употреблении, т. е. с учетом того, в каком контексте и в какой ситуации то или иное высказывание было произведено.

Прогресс в сфере компьютерных технологий влечет за собой прогресс в создании и совершенствовании средств автоматической обработки текста и, как результат, порождает новые парадигмы лингвистических исследований.

Дополняет ли корпус традиционную лингвистику или в какой-то степени заменяет ее? Очевидно, что дополняет. Но не просто дополняет экстенсионально, но и углубляет – по мере того как новая методология, вооруженная статистическим аппаратом, объединяется с традиционной лингвистической теорией. При этом положения и факты традиционной лингвистики во многом уточняются, а в ряде случаев появляются и новые лингвистические объекты, порожденные в недрах корпусной лингвистики и подлежащие изучению ее методами. В качестве примера можно привести понятие «коллострукция» (collostruction) [Stefanowich, Gries 2003]). Collostruction - термин-гибрид от слов collexeme (collocation + lexeme) и construction (лексемы, которые «притягиваются» специфическими конструкциями или конструкции, ассоциирующиеся с определенными лексемами). Авторы развивают семейство методов, направленных на измерение степени притяжения или «отталкивания» слов и конструкций (collostructional analysis).

Правомерно, видимо, говорить о новом интегрированном подходе, при котором корпусная лингвистика становится объединяющим началом, взаимодействуя с грамматикой, лексикологией, психолингвистикой, когнитивной лингвистикой. При этом наблюдается все более широкое распространение корпусной лингвистики – как материала и как метода – на всю сферу гуманитарных исследований, включая историю, социологию, литературоведение и т.д. Уже сегодня на базе корпусной методологии фактически сформировалась новая наука – культурометрия (culturomics) (в словаре Dictionary.com

этот термин определяется как «the study of human culture and cultural trends over time by means of quantitative analysis of words and phrases in a very large corpus of digitized texts» [Culturomics] - исследование культуры человечества, направлений её развития во времени посредством количественного анализа слов и словосочетаний в очень больших корпусах оцифрованных текстов). Социальный и культурный опыт человечества, зафиксированный в текстах, в лице корпусной лингвистики получает инструмент, позволяющий надеяться, что мы научимся автоматически извлекать из текстов знание. Хочется верить, что корпусная лингвистика вместе с традиционной лингвистикой, психолингвистикой и нейролингвистикой сформируют новую науку – интегрированную эмпирическую лингвистику – которая позволит глубже, чем до сих пор, понять фундаментальную природу языка.

Литература

Беликов В.И., Копылов Н.Ю., Пиперски А.Ч., Селегей В.П., Шаров С.А. Корпус как язык: от масштабируемости к дифференциальной полноте // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая – 2 июня 2013 г.). Вып. 12 (19). М.: Изд-во РГГУ, 2013. С. 84-95.

Беликов В.И. Yandex как лексикографический инструмент // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции «Диалог 2004», Москва,2004. С. 39-46.

Гришина Е.А. Мультимедийный русский корпус (МУРКО): проблемы аннотации // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 175–214.

Гришина Е.А., Савчук С.О. Корпус устных текстов в НКРЯ: состав и структура // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 129–149.

Захаров В.П. Веб-пространство как языковой корпус // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2005», Москва, 2005. С. 166-171.

Плунгян В.А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении, 2008, № 2. С. 7–20.

Шерстинова Т.Ю., Степанова С.Б., Рыко А. И. Система аннотирования в звуковом корпусе русского языка «Один речевой день» // Мат-лы XXXVIII международной филологической конференции. Секция: «Формальные методы анализа русской речи». СПб.: СПбГУ, 2009. С. 66-75.

Atkins, S., Clear, J., Ostler, N. (1992). Corpus Design Criteria. Literary and Linguistic Computing. Vol. 7. No. 1. 1–16.

Baker, P., McEnery, T., Hardie, A. (2006). A glossary of corpus linguistics. Edinburgh: Edinburgh University Press.

Baroni, M., Bernardini S., Ferraresi A., Zanchetta E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation. Vol. 43. No. 3. 209-226.

Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In: Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014.

Proceedings. LNCS 8655. Springer International Publishing Switzerland. 257-264.

Culturomics // Dictionary.com [Электронный ресурс] URL: http://dictionary.reference.com/browse/culturomics (дата обращения 11.03.2016)

Firth, J. R. (1957), A synopsis of linguistic theory 1930-1955, In: F. Palmer (Ed.), Selected Papers of J. R. Firth 1952-1959, London, Longman. 168-205.

Francis, N. W. (1991). Language Corpora B.C. Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82. Stockholm, 4.–6. August 1991. / Svartvik J. (ed.). 17–32.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel V. (2013).The TenTen Corpus Family. 7th International Corpus Linguistics Conference, Lancaster. Abstract Book. 137-139.

Kilgarriff, A. (2001). Web as corpus. P. Rayson, A. Wilson, T. McEncry, A. Hardic and S. Klioja (eds.) Proceedings of the Corpus Linguistics 2001 Conference, Lancaster (29 March-2 April 2001). Lancaster: UCREL. 342-344.

Kilgarriff, A., Grefenstette, G. (2003). Introduction to the Special Issue on Web as Corpus. Computational Linguistics. Vol. 29. No. 3. 333-347.

Leech,G. (1993). Corpus annotation schemes. Literary and Linguistic Computing. Vol. 8. No. 4. 275-281.

Renouf, A. (2003). WebCorp: providing a renewable data source for corpus linguists., S. Granger and S. Petch-Tyson (eds.) Extending the Scope of Corpus-based Research: New Applications, New Challenges. Amsterdam: Rodopi. 39–58.

Sharoff, S. (2006), Creating General-Purpose Corpora Using Automated Search Engine Queries. In WaCky! Working Papers on the Web as Corpus. Bologna: Gedit Edizioni. 63–98.

Sinclair, J. (1991). Corpus, Concordance, Collocation, Oxford University Press.

Sinclair, J. (2004). Trust the Text: Language, Corpus and Discourse. London: Routledge.

Stefanowitsch, A. and Gries, St. Th. (2003). Collostructions: investigating the interaction between words and constructions. International Journal of Corpus Linguistics. Vol. 8. No. 2: 209–243.

Stubbs, M. (1993). British traditions in text analysis: from Firth to Sinclair. M. Baker, F. Francis and E. Tognini-Bonelli (eds.) Text and Technology: In Honour of John Sinclair. Amsterdam: John Benjamins. 1–36.

Tognini-Bonelli, E. (2001). Corpus Linguistics at Work. Amsterdam: John Benjamins.

Xiao, R., Tono, Y. (2006). Corpus-based language studies: An advanced resource book. Taylor & Francis.

PROLEGOMENA TO CORPUS LINGUISTICS

Zakharov Victor Pavlovich
Associate Professor of the Mathematical Linguistics Departament
Saint-Petersburg State University
Saint-Petersburg, Universitetskaya emb.,7/9
e-mail: v.zakharov@spbu.ru

The paper introduces corpus linguistics as a new direction in linguistic science. The description of a subject and the main maintenance of corpus linguistics are given. A small historical digression is presented. The paper describes the concept ‘corpus’ and characterizes

the basic concepts of corpus linguistics: representativeness, balance, tagging, annotation, corpus manager, etc. The bases for typology and the main types of corpora are presented. The question of a subject of corpus linguistics, its place among other sections and directions in linguistics has risen and became more and more sharp as corpus technologies have been developing. Now corpus linguistics is broadening contact fields with theoretical linguistics and with different directions in it, therefore it is important to realize its place in the world of linguistics. This subject is discussed in the work, too. Prospects of the development, both within this direction, and together with other linguistic disciplines are given.

Keywords: corpus linguistics, corpus, tagging, annotation, representativeness, pragmatics, linguistic theory, statistical data

References

Atkins, S., Clear, J., Ostler, N. (1992). Corpus Design Criteria. Literary and Linguistic Computing. Vol. 7. No. 1. 1–16.

Baker, P., McEnery, T., Hardie, A. (2006). A glossary of corpus linguistics. Edinburgh: Edinburgh University Press.

Baroni, M., Bernardini S., Ferraresi A., Zanchetta E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation. Vol. 43. No. 3. 209-226.

Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S. (2013). Korpus kak yazyk: ot masshtabiruyemosti k differentsial'noy polnote [Corpus as language: from scalability to register variation]. Komp'juternaja lingvistika i intellektual'nye tekhnologii: po materialam ezhegodnoy mezhdunarodnoj konferentsii "Dialog 2013" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013"]. Vol. 12 (19), Moscow, RGGU, pp. 84–95.

Belikov V.I. Yandex kak lexikograficheskii instrument [Yandex as a lexicographical tool]. Komp'juternaja lingvistika i intellektual'nye tekhnologii: po materialam ezhegodnoy mezhdunarodnoj konferentsii "Dialog 2004" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2004"]. Vol. 3 (10), Moscow, pp. 39-46.

Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In: Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland. 257-264

Culturomics // Dictionary.com [Электронный ресурс] URL: http://dictionary.reference.com/browse/culturomics (дата обращения 11.03.2016)

Firth, J. R. (1957), A synopsis of linguistic theory 1930-1955, In: F. Palmer (Ed.), Selected Papers of J. R. Firth 1952-1959, London, Longman. 168-205.

Francis, N. W. (1991). Language Corpora B.C. Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82. Stockholm, 4.–6. August 1991. / Svartvik J. (ed.). 17–32.

Grishina E. A. Mul'timedijnyj russkij korpus (MURKO): problemy annotacii [Multimedia Russian corpus: annotation problems]. Nacional'nyj korpus russkogo jazyka: 2006–2008. Novye rezul'taty i perspektivy [National Russian corpus 2006-2008. New results and perspectives]. SPb, Nestor-Istorija, 2009, pp. 175–214.

Grishina E. A., Savchuk S. O. Korpus ustnyh tekstov v NKRJa: sostav i struktura [Oral

corpus in RNC: composition and structure]. Nacional'nyj korpus russkogo jazyka: 2006–2008. Novye rezul'taty i perspektivy. [National Russian corpus 2006-2008. New results and perspectives]. SPb, Nestor-Istorija, 2009, pp. 129–149.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel V. (2013).The TenTen Corpus Family. 7th International Corpus Linguistics Conference, Lancaster. Abstract Book. 137-139.

Kilgarriff, A. (2001). Web as corpus. P. Rayson, A. Wilson, T. McEncry, A. Hardic and S. Klioja (eds.) Proceedings of the Corpus Linguistics 2001 Conference, Lancaster (29 March-2 April 2001). Lancaster: UCREL. 342-344.

Kilgarriff, A., Grefenstette, G. (2003). Introduction to the Special Issue on Web as Corpus. Computational Linguistics. Vol. 29. No. 3. 333-347.

Leech,G. (1993). Corpus annotation schemes. Literary and Linguistic Computing. Vol. 8. No. 4. 275-281.

Plungyan V.A. Korpus kak instrument i kak ideologiya [Corpus as a tool and as ideology]. Russkii yazyk v nauchnom osvescenii [Russian in scientiphical light]. 2008, no. 2, pp. 7-20.

Renouf, A. (2003). WebCorp: providing a renewable data source for corpus linguists., S. Granger and S. Petch-Tyson (eds.) Extending the Scope of Corpus-based Research: New Applications, New Challenges. Amsterdam: Rodopi. 39–58.

Sharoff, S. (2006), Creating General-Purpose Corpora Using Automated Search Engine Queries. In WaCky! Working Papers on the Web as Corpus. Bologna: Gedit Edizioni. 63–98.

Sherstinova T. Ju., Stepanova S. B., Ryko A. I. Sistema annotirovanija v zvukovom korpuse russkogo jazyka «Odin rechevoj den'» [Annotation system in speech corpus "One spoken day"]. Mat-ly XXXVIII mezhdunarodnoj filologicheskoj konferencii [Proceedings of the XXXVIII. International Philologival Conference]. SPb, SPbGU, 2009, pp. 66-75.

Sinclair, J. (1991). Corpus, Concordance, Collocation, Oxford University Press.

Sinclair, J. (2004). Trust the Text: Language, Corpus and Discourse. London: Routledge.

Stefanowitsch, A. and Gries, St. Th. (2003). Collostructions: investigating the interaction between words and constructions. International Journal of Corpus Linguistics. Vol. 8. No. 2: 209–243.

Stubbs, M. (1993). British traditions in text analysis: from Firth to Sinclair. M. Baker, F. Francis and E. Tognini-Bonelli (eds.) Text and Technology: In Honour of John Sinclair. Amsterdam: John Benjamins. 1–36.

Tognini-Bonelli, E. (2001). Corpus Linguistics at Work. Amsterdam: John Benjamins.

Xiao, R., Tono, Y. (2006). Corpus-based language studies: An advanced resource book.

Zakharov V.P. Veb-prostranstvo kak yazykovoi korpus [Web as a language corpus]. Komp'juternaja lingvistika i intellektual'nye tekhnologii: po materialam ezhegodnoy mezhdunarodnoj konferentsii "Dialog 2005" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2005"]. Vol. 4 (11), Moscow, pp. 66-171.