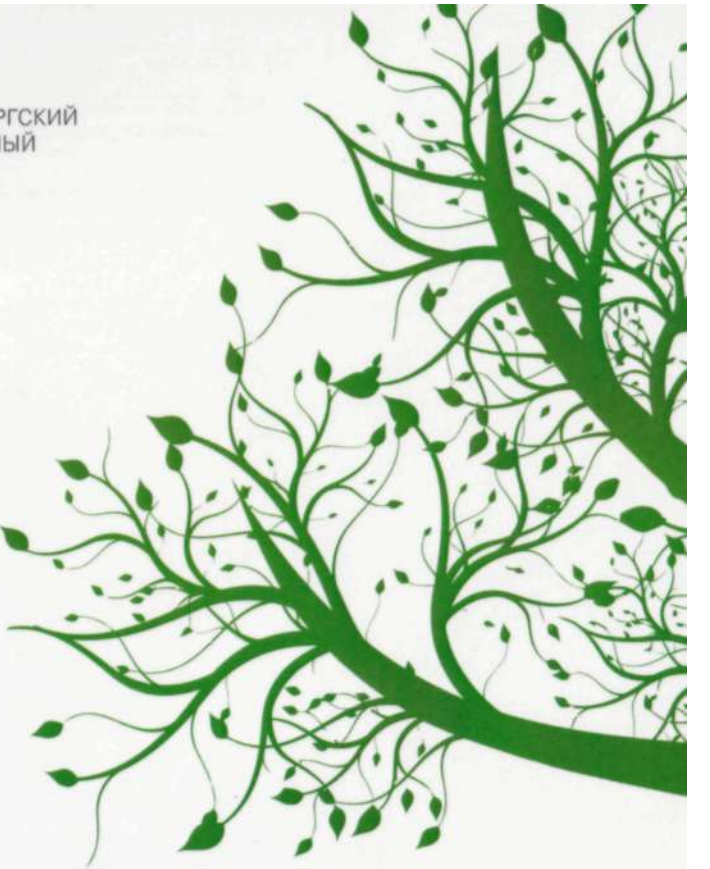
САНКТ-ПЕТЕРБУРГСКИЙ
ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ



В.П. Захаров, С.Ю. Богданова

КОРПУСНАЯ ЛИНГВИСТИКА

3-е издание

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

В. П. Захаров, С. Ю. Богданова

КОРПУСНАЯ ЛИНГВИСТИКА

3-е издание, переработанное

ИЗДАТЕЛЬСТВО САНКТ-ПЕТЕРБУРГСКОГО УНИВЕРСИТЕТА

УДК 81.32
ББК 81.1-923
З-38

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

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

Рекомендовано к публикации
Учебно-методической комиссией УГСН 45.00.00
Языкознание и литературоведение
Санкт-Петербургского государственного университета

Захаров В. П., Богданова С. Ю.

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

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

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

УДК 81.32
ББК 81.1-923

ISBN 978-5-288-05997-1

Оглавление

Предисловие к третьему изданию

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

По сравнению со вторым изданием главные изменения следующие:

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

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

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

Предисловие к первому и второму изданиям

Предлагаемый вашему вниманию учебник является своего рода обобщением многочисленных разрозненных материалов, опубликованных за последние два десятилетия в России и за рубежом. Данные материалы легли в основу лекционных курсов по дисциплине «Корпусная лингвистика», читаемых кандидатом филологических наук, доцентом Виктором Павловичем Захаровым в Санкт-Петербургском государственном университете и доктором филологических наук, профессором Светланой Юрьевной Богдановой в Иркутском государственном лингвистическом университете. Материал, представленный в учебном пособии, также может быть использован в курсах лекций по дисциплинам «Информационные и коммуникационные технологии в науке и образовании», «Основы прикладной лингвистики», «Компьютерные методы в лингвистических исследованиях» и др.

Цель учебника — познакомить студентов с концепциями корпусной лингвистики, помочь им освоить основы корпусных технологий, приобрести навыки работы с корпусами, определить место дисциплины и собственно корпусов в ряду информационно-лингвистических технологий.

Задачи учебного пособия:

- ознакомить студентов с новой парадигмой в лингвистических исследованиях;
- ознакомить студентов с историей корпусных исследований;
- ознакомить студентов с языковыми и программными средствами корпусной лингвистики;
- сформировать у студентов навыки работы с программными средствами и информационными ресурсами корпусной лингвистики;
- ознакомить студентов с конкретными лингвистическими исследованиями, основанными на корпусных данных.

Учебник состоит из трех частей. Первая часть — «Введение в корпусную лингвистику» — знакомит с основными понятиями и терминами корпусной лингвистики, историей ее становления как раздела языкознания, целями и задачами, типами существующих корпусов. Вторая часть — «Создание корпусов» — описывает в общих чертах технологические процессы, связанные с проектированием корпусов, отбором и обработкой языкового материала, способами разметки. Третья часть — «Использование корпусов» — включает три раздела. Раздел 3.1 «Корпусные менеджеры» посвящен описанию корпусных менеджеров, обеспечивающих поиск в корпусе. Раздел 3.2 «Обзор существующих корпусов различных типов» представляет собой обзор как зарубежных национальных корпусов, так и корпусов русского языка. Раздел 3.3 «Корпусные исследования» посвящен описанию конкретных исследований на базе корпусов разных типов, в нем приводятся результаты научных изысканий и дается их теоретическая интерпретация.

В первую очередь авторы хотят показать, как можно, базируясь на корпусах, работать с реальным языковым материалом быстрее и эффективнее. В этом разделе приведены примеры исследований лишь в нескольких областях лингвистики — лексикографии, грамматике и анализе дискурса. Безусловно, сфера применения корпусных данных в лингвистике значительно шире.

В приложении приведен краткий глоссарий терминов корпусной лингвистики.

Надеемся, что студенты направления «Лингвистика» заинтересуются использованием корпусов независимо от сферы их научных интересов, а каждый преподаватель найдет в учебнике то, о чем нужно говорить его аудитории.

Авторы выражают искреннюю благодарность заведующему кафедрой математической лингвистики СПбГУ Александру Сергеевичу Герду за критические замечания и рекомендации, сделанные в процессе подготовки учебника.

Часть 1

Введение в корпусную лингвистику

Глава 1. Основные понятия корпусной лингвистики

1.1. Определение корпусной лингвистики

Корпусная лингвистика — раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с применением компьютерных технологий. Под *лингвистическим, или языковым, корпусом текстов* (или обычно просто *корпусом текстов*) понимается большой, представленный в машиночитаемом формате, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач. Имея в виду круг задач (подчас достаточно широкий), для решения которых создается тот или иной корпус, можно говорить, что корпус всегда прагматически ориентирован.

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

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

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

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

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

Целесообразность создания и смысл использования корпусов определяются следующими предпосылками:

- достаточно большой (репрезентативный) и сбалансированный объем корпуса гарантирует типичность данных и обеспечивает полноту представления всего спектра языковых явлений;
- данные разного типа находятся в корпусе в своей естественной контекстной форме, что создает возможность их всестороннего и объективного изучения;
- однажды созданный и подготовленный массив данных может использоваться многократно, различными исследователями и в различных целях.

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

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

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

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

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

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

1.2. Предмет корпусной лингвистики

Сегодня корпусная лингвистика часто понимается как новая лингвистическая дисциплина, которая связана с изучением использования языка в реальной жизни с помощью компьютеров и электронных корпусов. Корпусная лингвистика имеет по крайней мере две черты, дающие ей основание претендовать на положение самостоятельной дисциплины: 1) характер используемого словесного материала, а именно размеченные тексты; 2) специфика инструментария.

Если такие разделы лингвистики, как синтаксис, семантика и социолингвистика, имеют целью описание или оценку языковой структуры или языкового использования, то корпусная лингвистика является более широким понятием, методологией, которую можно применить ко многим аспектам как языковых исследований, так и не только языковых. Корпусные методы лежат в основе новой дисциплины, которая получила название «культурометрия» (*culturomics*) и распространяется на все области гуманитарных исследований.

Корпусную лингвистику называют «пучком методов из разных областей лингвистических исследований» [Lüdeling, Kytö, 2008]. Как метод лингвистического анализа корпусная лингвистика связана также с контрастивными исследованиями, направленными на установление фактов общего и отдельного между языками, диалектами или

вариантами языка в ходе их сопоставительного изучения [Гвишиани, 2008]. Многие виды лингвистического анализа наилучшим образом развиваются на прочной и обширной базе эмпирических данных.

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

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

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

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

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

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

На наш взгляд, важным аспектом в определении корпусной лингвистики является то, что это не просто методология исследования языка — это наука, в недрах которой формируется сам объект исследования. Э. Финеган определяет корпусную лингвистику как деятельность, требующуюся для составления и использования корпуса и направленную на исследование естественного употребления языка [Finegan, 2004]. В этом определении подчеркивается созидательная направленность корпусной лингвистики. Ее двойственный характер (нацеленность как на создание, так и на использование корпусов текстов) обусловливается двойственным характером ее *объекта* — корпуса текстов, который, с одной стороны, представляет собой исходный речевой материал для корпусной лингвистики и для других лингвистических дисциплин; с другой стороны, сам является продуктом корпусной лингвистики.

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

1.3. Терминология корпусной лингвистики

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

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

гия складывалась и продолжает развиваться в недрах английского языка. Русские термины в основном представляют собой английские заимствования, некоторые из них, но в других значениях, давно существуют в русском языке. Так, русское слово «корпус» стало многозначным задолго до своего появления в качестве термина корпусной лингвистики. Употребление форм этого существительного в лингвистике является проблематичным, поскольку возможны варианты множественного числа «корпусы» и «корпуса». Для значения «массив», которое имеет место в случае языковых корпусов, именительный падеж множественного числа должен быть «ко́рпусы», и, соответственно, прилагательное «ко́рпусный» должно произноситься с ударением на первом слоге [Большой толковый словарь русского языка, 1998]. В то же время наблюдение над узусом специалистов пока свидетельствует в пользу форм «корпуса́», «корпусно́й», «корпусна́я», которые используются часто, так что можно, видимо, с осторожностью сказать, что в настоящее время этот вопрос остается открытым. Правила, регламентирующего употребление той или иной формы применительно к корпусной лингвистике, пока нет, хотя, как представляется, «победить» должен вариант «ко́рпусы». В данном учебнике авторы будут использовать именно этот вариант.

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

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

1.4. Направления в лингвистике, предвосхитившие появление корпусной лингвистики

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

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

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

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

Технологии, которые применяются в корпусной лингвистике, намного старше электронных компьютеров: многие из них коренятся в традиции конца XVIII — XIX в., когда лингвистика впервые была провозглашена реальной, или эмпирической, наукой. Из многочисленных областей лингвистических исследований, которые легли в основу корпусной лингвистики, здесь будут рассмотрены три. Использованные в этих трех областях технологии повлияли на

развитие современной корпусной лингвистики, и, наоборот, сейчас она существенно меняет «пейзаж» всей современной лингвистики, включая все нижеописанные направления [Lüdeling, Kytö, 2008].

1. Историческая лингвистика: изменения в языке и реконструкция (сравнительно-исторический метод). Одно из главных направлений, повлиявших на современную корпусную лингвистику, пришло из сравнительно-исторического языкознания. Это неудивительно, поскольку лингвисты, занимающиеся историческими исследованиями, всегда использовали тексты или собрания текстов как основные свидетельства. Многие технологии, развитые в XIX в., и в настоящее время используются для реконструкции более древних языков (праязыков) или установления связей между языками. В индоевропейской традиции изучение языковых изменений и попытки реконструкции зависели от ранних текстов или корпусов (исторических памятников). Я. Гримм и позднее младограмматики поддерживали утверждения об истории и грамматике языков цитатами из текстов. Младограмматики в своем манифесте провозгласили, что они провели исследование современного языка, зафиксированного в диалектах (а не только исследование древних текстов), и это также имело огромное значение.

Многие идеи, развиваемые с XIX в., были применены и затем развиты корпусной лингвистикой. Среди первых корпусов, доступных в электронном виде, были и исторические корпусы.

Появление огромного количества текстов, доступных в электронном формате, предоставило лингвистам возможность широко применять в лингвистическом анализе статистические методы, разрабатывать и развивать новые методы и модели исследований. Сегодня математически сложные модели языковых изменений могут быть построены на основе электронных корпусов.

2. Написание грамматик, составление словарей и обучение языку. Грамматисты XIX в. иллюстрировали свои утверждения примерами, взятыми из произведений признанных авторов. Например, Г. Пауль в своей немецкой грамматике использовал произведения классиков для иллюстрации каждого своего положения в области фонологии, морфологии и синтаксиса. Сегодня составители грамматик также используют корпусный подход, теперь корпусы включают не только классику, но и другие типы текстов и позволяют описать язык более адекватно. В частности, большой интерес сейчас вызывает грамматика устной речи.

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

Возьмем некоторые ранние примеры корпусного подхода из лексикографии. В середине XVIII в., когда С. Джонсон составлял толковый словарь английского языка (Dictionary of the English language, 1755), он выбирал из книг иллюстративные предложения, которые называл цитатами, чтобы показать на примерах, как слова использовались английскими авторами. Во время чтения С. Джонсон маркировал предложения, контекст которых делал значение слова особенно понятным. Его ассистенты выписывали отмеченные предложения на листы бумаги, и С. Джонсон распределял их для составления и иллюстрации словарных статей в словаре. Проект под руководством сэра Джеймса Муррея (Оксфордский словарь английского языка — OED) потребовал тысячи помощников и полвека для составления.

Многие словари мертвых языков давали цитаты из текстов, содержащие слово в контексте. В современной корпусной лингвистике этот метод параллелен по форме конкордансу KWIC (Key Word in Context), в котором искомое слово или конструкция выделяются в центре рабочего поля, а справа и слева отображается контекст. Компьютеры облегчили поиск и классификацию примеров, но идеи использования текстов из корпуса все еще очень схожи с теми, которых придерживались ранние лексикографы и филологи. И как мы уже писали выше, не только филологи.

[1] Термины «жанр» и «регистр» часто употребляются в литературе по корпусной лингвистике как синонимы, что, как представляется, зависит от предпочтений авторов. Тем не менее попытки «развести» эти термины неоднократно предпринимались (см., например: *Lee D.* Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle // Language Learning & Technology. 2001. Vol. 5, No. 3. P. 37–72 (http://llt.msu.edu/vol5num3/lee/default.html), где «жанры» определяются как группы текстов, собранных и скомпилированных для корпусов или корпусных исследований, которые понимаются как *категории* текстов, а «регистры» акцентируют внимание на параметрах ситуации языкового употребления и имеют естественную ассоциацию с определенными *лингвистическими чертами*). Это позволяет рассматривать устную речь (*spoken*) наравне с академической прозой (*academic*) и художественной литературой (*fiction*) как регистр..