Министерство образования и науки РФ

ФГБОУ ВО «Тверской государственный университет»

Факультет иностранных языков и международной коммуникации

Направление «Лингвистика»

Кафедра английского языка

Реферат

по дисциплине: «Информационные технологии в лингвистике»

на тему: «Основные виды корпусов»

|  |  |
| --- | --- |
|  | Выполнила:  студентка I курса факультета ИЯ и МК  группа 13  Мустафаева Тамиля |
|  |  |
|  | Проверила:  к.ф.н., доцент кафедры английского языка  Е.М. Масленикова |

Тверь

2019

**Оглавление**

|  |  |
| --- | --- |
| Введение………...……………………………………………………………….. | 3 |
| Корпусная лингвистика.……………………………………………………….. | 3 |
| Применение корпусов…………...……………………………………………… | 4 |
| История………………………………………………………………………….. | 6 |
| Типы корпусов…………………………………………………………………... | 8 |
| Корпусы в сети интернет……………………………………………………….. | 10 |
| Заключение………………………………………………………………………. | 13 |
| Список использованной литературы…………………………………………… | 14 |

**Введение**

За последние два десятилетия корпуса стали использоваться не только в лингвистических исследованиях, но и в преподавании и изучении языка, хотя, по всей видимости, создатели корпусов даже не предполагали такого их использования. На сегодняшний день существует значительное число работ, полностью основанных на корпусах (например, словари и грамматики) и доступных учителям и студентам, а также разработки по применению конкордансов и упражнений, основанных на работе с корпусами, которые могут быть использованы на уроках. Увеличивается количество печатных материалов и ресурсов для обучения иностранным языкам, отмеченных значком corpus-based («создано с использованием корпусных данных»). Достаточно сложно найти современный словарь известного издательства, который не был бы составлен на основе лингвистического корпуса.

Корпусные данные заставили по-новому взглянуть на наши представления о языке и пересмотреть подход к составлению грамматических и лексических справочников, учебников и других материалов для желающих овладеть иностранным языком. Таким образом, вклад корпусной лингвистики в описание языка, а также в методику преподавания иностранных языков представляется трудно оспоримым. Корпусная лингвистика привносит инновационные изменения в научные методы и технологии и, возможно, предвосхищает появление еще более основательных перемен, которые столкнутся с нашими традиционными представлениями об образовании, о роли преподавателя, о культурном контексте оказания образовательных услуг.(1)

**Корпусная лингвистика**

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

В приведенных определениях подчеркиваются основные черты современного корпуса текстов – цель («логическая идея»), 8 машиночитаемый формат, репрезентативность как результат особой процедуры отбора, наличие металингвистической информации. Стандартизованное представление словесного материала на машинном носителе позволяет применять стандартные программы его обработки.(2)

**Применение корпусов**

Корпусная лингвистика — раздел прикладной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов при помощи компьютеров.

Исходя из такого определения можно констатировать, что корпусная лингвистика включает два аспекта:

1. создание корпусов текстов с автоматическими инструментами их использования;
2. разработка способов экспериментальных исследований различных уровней языка на базе корпусов разных типов.

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

Кроме проведения научных исследований корпусы могут использоваться :

1. *в лексикографии* для создания словарей, определения значения многозначных слов и т.д.;
2. *в грамматике* для определения частоты морфем, типов словосочетаний и предложений и т.д.;
3. *в лингвистике* текста для дифференциации типов текста, выявления связей внутри абзаца и между абзацами и т.д.;
4. *в автоматическом переводе* текстов для поиска контекстов слов, имеющих несколько переводных эквивалентов, поиска переводных эквивалентов в параллельных текстах и т.д.;
5. *в учебных целях* для выбора цитат, фрагментов произведений, примеров для организации учебных занятий, создания учебных пособий и т.д.
6. *в тестировании* программ автоматического анализа и синтеза речи и т.д.

Центральное понятие корпусной лингвистики — *лингвистический корпус* — определяется как совокупность специально отобранных текстов, размеченных по различным лингвистическим параметрам и обеспеченных системой поиска. таким образом, корпус можно кратко охарактеризовать следующим образом:

Корпус = тексты + их разметка.

В более широком смысле корпусом считается любое собрание текстов. В этой трактовке выделяются размеченные (аннотированные) и неразмеченные корпусы текстов. В качестве подобных неразмеченных корпусов можно рассматривать существующие электронные коллекции текстов: виртуальные библиотеки, архивы электронных версий периодических изданий или новостных лент, которые оказываются достаточными для некоторых исследовательских и учебных целей. но использование неразмеченных собраний текстов, имеющих инструменты поиска, повышает долю информации, которая может оказаться нерелевантной для исследователя, что значительно затрудняет работу с таким источником. В связи с этим предметом корпусной лингвистики являются преимущественно размеченные корпусы текстов.(3)

**Из истории**

Лингвисты собрали первые корпусы компьютеризированных текстов в 1960-е годы. Первый компьютеризированный корпус – Брауновский корпус (The Brown Corpus1) – включает 500 текстов из американских книг, газет, журналов, впервые опубликованных в США в 1961 году. Каждый текст в Брауновском корпусе имеет длину 2000 слов (имеется в виду словоупотреблений – tokens), и все собрание включает 1 млн. слов (500 текстов по 2000 слов в каждом). Авторы корпуса У. Френсис (W. Francis) и Г. Кучера (H. Kucera) сопроводили его большим количеством материалов первичной статистической обработки: частотным и алфавитно-частотным словарем, разнообразными статистическими распределениями.

Цель создания Брауновского корпуса – обеспечить системное изучение отдельных жанров письменного английского языка и сравнение жанров. Его появление вызвало всеобщий интерес и оживленные дискуссии. В первую очередь, они коснулись принципов отбора текстов и состава, потенциально решаемых на таком корпусе задач. С одной стороны, он строился на основе статистических процедур; с другой стороны, статистика применялась в сочетании с волевыми решениями авторов корпуса, базирующимися на профессиональной интуиции. Для достижения максимальной объективности этого сложного процесса требовалось построение максимально формализованных, прозрачных для проверки и контроля процедур.

Позднее европейские исследователи составили корпус текстов, впервые опубликованных в Великобритании в 1961 году, следуя тем же принципам: 15 жанров (регистров), 500 текстов по 2000 слов (словоупотреблений). Он включал 1 млн. слов британского варианта английского языка, и его назвали корпусом Ланкастер-Осло-Берген (The Lancaster-Oslo-Bergen Corpus, по названиям британского и двух норвежских университетов, или кратко LOB). Сбалансированные корпусы типа Брауновского очень важны для исследователей, чьи интересы лежат в области лингвистики и которые хотят использовать корпус в целях лингвистического описания и анализа. Итак, два самых ранних больших корпуса – это корпусы письменной речи американского и британского вариантов английского языка. Оба корпуса остаются полезными и сейчас, на них основываются многочисленные исследования английского языка.

За десятилетия, прошедшие с момента создания этих корпусов, компьютеры стали дешевле и гораздо мощнее, кроме того, недорогие и надежные сканеры сделали необязательным набор текстов на компьютере с помощью клавиатуры. Эти изобретения облегчили процесс создания корпусов, и последние из них содержат уже миллиарды слов (словоупотреблений). К 1990 году уже было зафиксировано более 600 компьютерных корпусов. По годам составления они были распределены примерно следующим образом: - 1965 10 1966-1970 20 1971-1975 30 1976-1980 80 1981-1985 160 1986-1990 320 Очевидно, что в последующие годы количество и многообразие создаваемых корпусов шли по нарастающей. Среди современных корпусов английского языка (как британского, так и американского варианта) наиболее известны Британский национальный корпус (British National Corpus – BNC), Международный корпус английского языка (International Corpus of English – ICE), лингвистический Банк английского языка (Bank of English), Корпус современного американского английского (Corpus of Contemporary American English – COCA) и др. В настоящее время корпусы созданы для многих языков мира .

В первой половине 1990-х годов корпусная лингвистика окончательно сформировалась как отдельное направление науки о 17 языке. «Корпусная лингвистика достигла зрелости» – так Я. Свартвик озаглавил в 1992 году предисловие к материалам первого Нобелевского симпозиума по корпусной лингвистике. Корпусная лингвистика тесно взаимодействует с компьютерной лингвистикой, используя ее достижения и, в свою очередь, обогащая ее.(4)

**Типы корпусов**

Несмотря на разнообразие корпусов, можно выделить два основных способа деления корпусов на классы: 1) 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.); 2) разделение корпусов по типу лингвистической разметки. Несмотря на наличие множества типов разметки, большинство реально существующих корпусов относится к корпусам морфологического либо синтаксического типа (последние в англоязычной литературе называют treebanks, что можно перевести как «банки синтаксических структур»). При этом следует подчеркнуть, что корпус с синтаксической разметкой явно или неявно включает в себя и морфологические характеристики лексических единиц.

Вообще же существует большое число разных типов корпусов. Их разнообразие определяется многообразием исследовательских и прикладных задач, для решения которых они создаются, и различными основаниями для классификации. В зависимости от поставленных целей и классифицирующих признаков, можно выделить различные типы корпусов (см. таблицу).(5)

***Классификация корпусов***

|  |  |
| --- | --- |
| Признак | Типы корпусов |
| Тип данных | Письменные  Речевые  Смешанные |
| Язык текстов | Русский  Английский и т.д. |
| «Параллельность» | Одноязычные  Двуязычные  Многоязычные |
| «Литературность», специфичность | Литературные  Диалектные  Разговорные  Терминологические  Смешанные |
| Жанр | Литературные  Фольклорные  Драматургические  Публицистические |
| Доступность | Свободно доступные  Коммерческие  Закрытые |
| Назначение | Исследовательские  Иллюстративные |
| Динамичность | Динамические (мониторные)  Статические |
| Разметка | Размеченные  Неразмеченные |
| Характер разметки | Морфологические  Синтаксические  Семантические  Просодические и т.д. |
| Объем текстов | Полнотекстовые «Фрагментнотекстовые» |
| Хронологический аспект | Синхронические Диахронические |
| «Общность» | Общие  Одного писателя |
| Структура | Центральные и архивные  Ядерные и периферийные |

**Корпусы в сети Интернет**

Приведем сетевые адреса и краткие сведения о некоторых корпусах. В Интернете можно получить доступ и найти списки самых различных корпусов — см., например, D. Lee. Bookmarks for Corpus-based Linguists (<http://devoted.to/corpora> ), веб-страницы М. Барбера (Manuel Barbera) (<http://www.bmanuel.org/index.html> ) или М. Барлоу (Michael Barlow) (<http://www.athel.com/corpus.html> ), сайт Language and Speech Resources (<http://www.elsnet.org/resources.html> ) и др.

|  |  |
| --- | --- |
| Национальный корпус русского языка  [**http://ruscorpora.ru**](http://ruscorpora.ru) | 70 млн слов[[1]](about:blank) |
| Компьютерный корпус текстов русских газет конца ХХ-го века [**http://www.philol.msu.ru/~lex/c+orpus**](http://www.philol.msu.ru/~lex/c+orpus) | 200 тыс. слов Система поиска по корпусу временно недоступна |
| Корпус русского языка ХАНКО (Хельсинский университет) [**http://www.ling.helsinki.fi/projects/hanco/**](http://www.ling.helsinki.fi/projects/hanco/) | 100 тыс. слов  Ручная морфологическая разметка |
| Корпуса русских текстов на сайте Университета в Лидсе, Великобритания [**http://corpus.leeds.ac.uk**](http://corpus.leeds.ac.uk) |  |
| Русские корпуса Тюбингенского Университета [**http://www.sfb441.uni-tuebingen.de/b1/en/korpora.html**](http://www.sfb441.uni-tuebingen.de/b1/en/korpora.html) |  |
| Словарь-корпус языка А.С. Грибоедова[**http://www.inforeg.ru/electron/concord/concord.htm**](http://www.inforeg.ru/electron/concord/concord.htm) | 120 тыс. слов |
| Уппсальский корпус русских текстов Доступен для поиска на сайте [http://www.sfb441.uni-**tuebingen.de/b1/en/korpora.html**](http://www.sfb441.uni-tuebingen.de/b1/en/korpora.html) | 1 млн слов  600 текстов (публицистика 1985-1989; литературные произведения 1960-1988). |
| Банк английского языка (Bank of English)  [**http://www.collins.co.uk/books.aspx?group=153**](http://www.collins.co.uk/books.aspx?group=153)  Свободный доступ: [**http://www.collins.co.uk/Corpus/CorpusSearch.aspx**](http://www.collins.co.uk/Corpus/CorpusSearch.aspx) | 524 млн слов, 56 млн в свободном доступе (The Collins Wordbanks *Online*English corpus: 36 млн – брит. англ., 10 млн – амер. англ., 10 млн – брит. разговорн. англ.) |
| Британский национальный корпус  [**http://www.natcorp.ox.ac.uk/**](http://www.natcorp.ox.ac.uk/) или **http://sara.natcorp.ox.ac.ux**/ | 100 млн слов Корпусные менеджеры SARA и XAIRA (http://www.xaira.org) |
| Венгерский национальный корпус  [**http://corpus.nytud.hu/mnsz**/](http://corpus.nytud.hu/mnsz/) | 100 млн слов |
| Корпус испанского языка (исторический) [**http://www.corpusdelespanol.org**/](http://www.corpusdelespanol.org/) | 100 млн слов, тексты 13–20 вв. Создан в Иллинойском университете, США |
| Корпус латинских текстов «Персей» [**http://www.perseus.tufts.edu**](http://www.perseus.tufts.edu) |  |
| Корпус современного датского языка  [**http://www.korpus2000.dk/**](http://www.korpus2000.dk/) | 50 млн слов  Тексты 1998–2002 гг. |
| Корпус современного итальянского языка CORIS/CODIS  [**http://www.cilta.unibo.it/ricerca.htm**](http://www.cilta.unibo.it/ricerca.htm) | 100 млн Слов |
| Корпус современного китайского языка (LIVAC Synchronous Corpus) [**http://www.rcl.cityu.edu.hk/livac/**](http://www.rcl.cityu.edu.hk/livac/) | 720 млн слов  (150 млн иероглифов) |
| Мангеймский корпус немецкого языка (Institut für Deutsche Sprache, Mannheim, Germany) [**http://corpora.ids-mannheim.de/~cosmas/**](http://corpora.ids-mannheim.de/~cosmas/) | 1610 млн слов  Корпусный менеджер COSMAS |
| Национальный корпус словенского языка  [**http://www.fida.net/eng/**](http://www.fida.net/eng/) | Более 100 млн слов |
| Польский национальный корпус [**http://korpus.ia.uni.lodz.pl/**](http://korpus.ia.uni.lodz.pl/) | 93 млн слов |
| Словацкий национальный корпус  [**http://korpus.juls.savba.sk**](http://korpus.juls.savba.sk) | 180 млн слов Используется корпусный менеджер Manatee/Bonito. |
| Хорватский национальный корпус  [**http://www.hnk.ffzg.hr/**](http://www.hnk.ffzg.hr/) | 53 млн слов  Корпусный менеджер  Manatee/Bonito. |
| Чешский национальный корпус  [**http://ucnk.ff.cuni.cz**](http://ucnk.ff.cuni.cz) | 100 млн слов + 100 млн нового корпуса современной лексики Корпусный менеджер  Manatee/Bonito. |

**Заключение**

Таким образом, корпус – это представленный в электронном виде, как правило, размеченный для анализа в лингвистических целях, обеспеченный сравнительно простой в использовании поисковой системой репрезентативный массив неотредактированных текстов, представляющих как можно большее количество «вариантов» языка. В период зарождения корпусной лингвистики вопросов компьютеризации данного направления не ставилось, и «исследователи указывали на возможность пренебречь языковой вариативностью, т. е. территориальной, социальной, профессиональной, возрастной, гендерной, индивидуальной и тому подобной дифференциацией языка» [Плунгян, 2006. С. 76–77].

Сегодня же, пренебрегая ею, мы сознательно ограничиваем себя различными рамками при изучении текстов определенного языка, что ставит под вопрос объективность подобного рода исследования. С появлением электронных корпусов многообразие форм существования языка стало более наглядным, возможности исследования языковых данных расширились. Современный лингвистический корпус содержит сотни миллионов словоупотреблений, а то, что с помощью электронного корпуса результаты примеров словоупотреблений можно получить за считанные доли секунд, существенно упрощает задачу лингвистам. Представленная типология корпусов, не претендуя на всеохватность, показывает нам существующее многообразие корпусов текстов и позволяет сориентироваться в нем для последующего проведения научного исследования.(6)

**Список использованной литературы**

1. «Корпусная лингвистика и обучение языку» https://revolution.allbest.ru/languages/00692256\_0.html
2. В. П. Захаров, С. Ю. Богданова «Корпусная лингвистика» с. 7; 2011 г.
3. Л.Ю. Щипицина «Информационные технологии в Лингвистике» с 57-59; 2013 г.
4. В. П. Захаров, С. Ю. Богданова «Корпусная лингвистика» с.14- 17; 2011 г.
5. В. П. Захаров «Корпусная лингвистика 2005 г.
6. Корпусы и корпусная лингвистика. Основные понятия-Корпусная лингвистика https://www.myfilology.ru/177/korpusy-i-korpusnaya-lingvistika-osnovnye-ponyatiya/#
7. Плунгян В. А. «Интегрум» и Национальный корпус русского языка в лингвистических исследованиях // Integrum: точные методы и гуманитарные науки. М., 2006. С. 76–77