Министерство Образования Российской Федерации

ФБГОУ ВПО «Тверской государственный университет»

Факультет иностранных языков и международной коммуникации

Направление «Лингвистика»

РЕФЕРАТ

«Типы и разметки корпусов»

|  |  |
| --- | --- |
|  | Выполнила:  студентка второго курса 21 группы, ЗФО  Горшанова Александра Витальевна |
|  |  |
|  | Проверила:  кфн, доцент кафедры английского языка Масленникова Елена Михайловна |

Тверь, 2020 год

СОДЕРЖАНИЕ

[Введение 3](#_Toc58311297)

[Корпусная лингвистика 4](#_Toc58311298)

[Принципы корпусной лингвистики 7](#_Toc58311299)

[Корпус текстов и корпус данных 9](#_Toc58311300)

[Типы корпусов 10](#_Toc58311301)

[Требования к корпусам 12](#_Toc58311302)

[Разметка корпусов 13](#_Toc58311303)

[Национальный корпус русского языка 15](#_Toc58311304)

[Литература 19](#_Toc58311305)

# Введение

Информационные технологии в настоящее время являются неотъемлемой частью любой сферы профессиональной деятельности, в том числе лингвистики. И если когда-то использование компьютеров и соответствующих программ в лингвистических исследованиях, переводе и в обучении языку не являлось обязательным, то сегодня уже со студенческой скамьи будущим преподавателям иностранных языков, переводчикам и лингвистам-исследователям необходимы компетенции, связанные с использованием информационных технологий в своей профессиональной сфере деятельности.

Корпусная лингвистика

Корпусная лингвистика - это раздел прикладной лингвистики, занимающийся разработкой общих принципов построения и использованием лингвистических корпусов (корпусов текстов).

Одним из основных источников языкового материала, необходимого для проведения лингвистических и филологических исследований, является текст, письменный или устный. На основе анализа множества текстов можно сделать вывод об интересующем исследователя языковом явлении, например, о поведении грамматической конструкции, использовании выразительных средств в языке и т.п. Развитие вычислительной техники способствовало тому, что большое количество текстов стало доступно в электронном виде. Для того чтобы можно было работать с такими объемами текстов, извлекая из них нужную информацию, во всем мире стали создаваться лингвистические корпусы, т.е. коллекции текстов, специально отобранных, размеченных по различным лингвистическим параметрам и обеспеченных системой поиска. Увеличение объема материала исследования потребовало применения новых методов анализа лингвистических данных, включая их статистическую обработку.

Таким образом, корпусная лингвистика включает два аспекта: во-первых, создание и разметка (аннотирование) корпусов текстов и разработка средств поиска по ним и, во-вторых, собственно лингвистический аспект - экспериментальные исследования на базе корпусов.

Возможность пользоваться обширными электронными ресурсами значительно облегчила процесс сбора материала в лингвистических исследованиях. Однако такая доступность языковых данных радикально изменила требования к доказательной базе лингвистических исследований: в наше время недостаточно привести несколько примеров в подтверждение выдвигаемой концепции, необходимо представить достаточно полную выборку данных того или иного языка. Однако работать с мегабайтами «сырых» текстов не менее трудно, чем искать примеры вручную.

Необходимость и возможность обработки большого массива текстов для извлечения из них лингвистических, литературоведческих и других данных обусловили бурный рост электронных ресурсов, включая лингвистически аннотированные корпуса текстов.

Использование корпусов текстов дает возможность наблюдать поведение интересующих исследователя языковых единиц (слов, словосочетаний, грамматических категорий, синтаксических конструкций и т.д.) в естественной языковой среде, т.е. в реально существующих, а не искусственно сконструированных контекстах.

Кроме того, корпусные исследования позволяют, используя статистические методы, сформулировать, подтвердить или опровергнуть некоторую гипотезу о том или ином языковом явлении на большом объеме материала.

При этом если исследователь пользуется уже существующим корпусом, он полностью минует долгий и трудоемкий этап сбора материала (опрос информантов, работа со словарными картотеками или письменными текстами и т.д.).

Для некоторых целей оказывается достаточным использование в качестве корпуса уже существующих электронных коллекций текстов, таких как виртуальные библиотеки, архивы электронных версий периодических изданий или новостных лент. Гигантским собранием текстов является весь Интернет.

Одним из приоритетных направлений корпусной лингвистики является создание и расширение универсальных национальных корпусов (корпусов того или иного естественного языка), представительных по отношению ко всему языку, которые могут служить для исследования самых разнообразных явлений этого языка. Большинство языков мира уже имеют свои национальные корпуса. Общепризнанным образцом является, в частности, Британский национальный корпус (BNC). Среди корпусов славянских языков выделяется Чешский национальный корпус, созданный в Карловом университете Праги. Для русского языка таким представительным корпусом является Национальный корпус русского языка (НКРЯ).

По мере создания все большего количества корпусов и развития различных технологий обработки текстовой информации вырабатывались некоторые универсальные принципы и языки разметки лингвистической информации в корпусе, создавались специальные инструменты для работы с корпусом на любом языке так называемые корпусные менеджеры. Выработка универсальных стандартов и технологий сделало возможным создавать большие представительные корпуса за очень короткий срок.

Таким образом, предметом корпусной лингвистики являются «теоретические основы и практические механизмы создания и использования представительных массивов языковых данных, предназначенных для лингвистических исследований в интересах широкого круга пользователей».

# Принципы корпусной лингвистики

Корпусная лингвистика занимается разработкой общих принципов построения и использования лингвистических корпусов данных с использованием компьютерных технологий. Она использует три приема:

* автоматическое извлечение сведений о языке из корпусов;
* обработка информации;
* проверка и интерпретация обработанных данных.

Первые две ступени полностью алгоритмизированы, третья является спорной.

В основе корпусной лингвистике лежит то, что язык — это полностью социальное явление, и его можно описать данными основанными на опыте, т.е. в речевом акте. Это влечет за собой то, что мы не знаем, а иногда и стараемся не знать, как говорящий или слушающий понимает слова, предложения или тексты, которые он говорит или слышит. Как социальный феномен язык проявляет себя в текстах, которые можно записать, описать и проанализировать. Внутренние, немые тексты также являются текстами, но их нельзя пронаблюдать и, следовательно, они не являются социальным явлением. Большинство текстов встречаются в виде речевых актов, т.е. как взаимодействие между членами языкового общества.

Долгое время люди не могли записывать и тем самым фиксировать речевые акты, но в наше время при наличии разнообразной аппаратуры это сделать довольно просто, поэтому количество записанных речевых актов значительно выросло. При этом возникает проблема обработки всей информации. Следовательно, этот поток информации должен быть сжато выражен, к примеру, в виде отдельного явления, который можно описать. Параметрами могут быть язык, отрезок времени, район, внутренние или внешние свойства текстов и т.д.

Уникальная черта естественного языка состоит в том, что он динамичен. Существуют правила речевого акта, которые можно обсуждать, опровергать и отвергать, а также могут быть распространены новые правила. Также предметом обсуждения являются значения, несмотря на все попытки лексикографов стандартизировать их. Корпусная лингвистика описывает отдельные естественные языки, индивидуальные лингвистические качества. Поскольку отдельные элементы языка и речи характеризуются некоторой частотой, а следовательно, и вероятностью, произведения речи очевидно могут анализироваться и исследоваться с помощью методов математической статистики. Ученые-лингвисты не могли не заметить статистических характеристик элементов речи, поэтому родилось направление лингвостатистики.

Следовательно, корпусная лингвистика смотрит на полную картину, чтобы узнать соглашения языкового общества. Для корпусной лингвистики, язык – это виртуальное функционирующее явление, которое может быть зафиксировано и представлено в виде модели человеческой речи в текстах.

# Корпус текстов и корпус данных

Прежде чем говорить про корпус текстов нужно понять, что такое корпус данных. Корпус данных представляет собой сформированную по определенным правилам выборку данных из области реализации языковой системы, которая содержит феномены, подлежащие лингвистическому описанию. Корпус данных имеет только одно измерение – речевое, поскольку сам по себе он не обладает потенцией производства своих составляющих. Последнее, однако, не означает, что корпус данных не может использоваться для реконструкции языка как системы. Наоборот, это одна из главных задач лингвистического исследования корпуса. Перед нами одно из глобальных противоречий, свойственное любому продукту языковой системы – от звука до текста. Лингвисту приходиться по отдельным результатам деятельности языка делать выводы о функционировании языка как целого, как системы.

Корпус текстов – это вид корпуса данных, единицами которого являются тексты или их достаточно значительные фрагменты, включающие, например, какие-то отрывки текстов данной проблемной области.

# Типы корпусов

В зависимости от поставленной цели выделяют несколько типов корпусов текстов:

1. по форме хранения:
   1. звуковой форме;
   2. письменные;
   3. смешанные;
2. по языку представления текстов:
   1. одноязычные;
   2. многоязычные;
3. по жанровой принадлежности:
   1. литературные;
   2. диалектные;
   3. разговорные;
   4. публицистические;
   5. смешанные;
4. по способам доступа:
   1. свободно доступные;
   2. коммерческие;
   3. закрытые;
5. по назначению:
   1. исследовательские;
   2. иллюстративные;
6. по динамичности:
   1. динамические (мониторные);
   2. статические;
7. по наличию дополнительной информации:
   1. аннотированные (размеченные);
   2. неразмеченные

# Требования к корпусам

При конструировании и применении корпусов единой методики для всех языков нет, так как различаются языки, традиции, технологические процессы. Но основные требования В.В. Рыков выделяет следующие:

1. Кто пользователь корпуса? (индивид, группа, лингвистическое общество).
2. Какова логическая идея, которая положена в основу корпуса?
3. С каким объёмом данных мы будем работать при составлении корпуса? На сколько это необходимо и реалистично?
4. Используем отрывки из текстов, полные тексты или то и другое.
5. Процедура отбора текстов в корпусе. Для разных целей по-разному:
   1. обследование речевого материала,
   2. сканирование текстов,
   3. окончательное формирование, составление корпуса.
6. Стандартизированное представление корпуса на уровне отраслевых стандартов, т.е., представление всего корпуса как продукта:
   1. аннотация всего текста в целом
   2. унифицированное представление словесного материала текста.
7. Аннотирование, индексирование словесного материала текста.

# Разметка корпусов

Разметка (англ. annotation) – содержащаяся в корпусе лингвистическая или иная информация, приписанная тем или иным отрезкам текста (так, информация о времени создания и жанре приписывается целому тексту, информация о синтаксической структуре – предложению, информация о лексеме и грамматических характеристиках – слову), а равно и процесс, добавляющий к исходному тексту эту информацию.

Разметка (tagging, annotation) заключается в приписывании текстам и их компонентам специальных меток (tag, tags): внешних, экстралингвистических (сведения об авторе и сведения о тексте: автор, название, год и место издания, жанр, тематика; сведения об авторе могут включать не только его имя, но также возраст, пол, годы жизни и многое другое. Это кодирование информации имеет название метаразметка), структурных (глава, абзац, предложение, словоформа) и собственно лингвистических, описывающих лексические, грамматические и прочие характеристики элементов текста. Набор этих метаданных во многом определяет возможности, предоставляемые корпусами исследователям.

* морфологическая разметка
  + в иностранной терминологии употребляется термин part-of-speech tagging (POS-tagging), дословно – частеречная разметка
  + выделение аффиксов, сложных слов и т.п.
  + основной тип разметки
  + морфологический анализ рассматривается как основа для дальнейших форм анализа – синтаксического и семантического
* синтаксическая разметка
  + результат синтаксического анализа, или парсинга (англ. parsing), выполняемого на основе данных морфологического анализа
  + синтаксические связи между лексическими единицами и различные синтаксические конструкции
  + характер синтаксической связи, тип предложения, член предложения
* семантическая разметка
  + семантические категории, к которым относится данное слово или словосочетание
  + более узкие подкатегории, специфицирующие его значение
  + снятие семантической омонимии, разрешение анафоры и кореферентности, фиксирование информационной структуры
* анафорическая разметка
  + фиксирует референтные связи, например, местоименные
  + указание для каждой словоформы из текста ее исходной формы
* просодическая разметка (дискурсная)
  + ударение и интонация
  + обозначение пауз, повторов, оговорок, и т.д.
  + реплики, коммуникативные акты и т.п.

# Национальный корпус русского языка

Национальный корпус предназначен в первую очередь для обеспечения научных исследований лексики и грамматики языка, а также тонких, но непрерывных процессов языковых изменений, происходящих в языке на протяжении сравнительно небольших периодов — от одного до двух столетий. Другая задача корпуса — предоставление всевозможных справок, относящихся к указанным областям (лексика, грамматика, акцентология, история языка). Современные компьютерные технологии многократно упрощают и ускоряют процедуры лингвистической обработки больших массивов текстов.

С этой точки зрения основными потребителями национальных корпусов являются, конечно, исследователи-лингвисты самого разного профиля. Однако круг пользователей корпуса вовсе не ограничивается профессиональными исследователями языка. Надежные статистические данные о языке определенной эпохи или определенного автора могут интересовать литературоведов, историков и представителей многих других областей гуманитарного знания. Важное значение национальные корпуса имеют и для преподавания языка в качестве родного или иностранного; всё больше учебников и учебных программ в наше время оказываются ориентированы на корпус. Быстро и эффективно проверить с помощью корпуса особенности употребления незнакомого слова или грамматической формы у авторитетных авторов сможет и иностранец, и школьник, и учитель, и журналист, и редактор, и писатель. Таким образом, национальный корпус обращен ко всем, кто в силу профессии, по необходимости или из простой любознательности ищет ответ на вопросы об устройстве и функционировании языка, то есть фактически к большинству образованных носителей этого языка и ко всем, изучающим его в качестве иностранного.

Национальный корпус русского языка охватывает прежде всего период от середины XVIII до начала XXI века: этот период представляет как язык предшествующих эпох, так и современный, в разных социолингвистических вариантах — литературном, разговорном, просторечном, отчасти диалектном. В корпус включаются оригинальные (непереводные) произведения художественной литературы (проза и драматургия, в дальнейшем также поэзия), имеющие культурную значимость, а также представляющие интерес с точки зрения языка. Но Национальный корпус ни в коей мере не является только корпусом языка художественной литературы. Помимо художественных текстов, в корпус в большом количестве включаются и другие образцы письменного (а для современного этапа — и устного) языка: мемуары, эссеистика, публицистика, научно-популярная и научная литература, публичные выступления, частная переписка, дневники, документы и т. п.

Национальный корпус русского языка в настоящее время включает следующие подкорпуса:

* основной корпус, в который входят прозаические (включая драматургию) письменные тексты XVIII — начала XXI века;
* синтаксический (глубоко аннотированный) корпус, в котором для каждого предложения построена полная морфологическая и синтаксическая структура (дерево зависимостей);
* газетный корпус (корпус современных СМИ), в котором представлены статьи из средств массовой информации 1990-2000-х годов;
* параллельные корпуса, в которых можно найти все переводы для определенного слова или словосочетания на русский язык или с русского языка. В настоящее время для поиска доступны англо-русский, русско-английский, немецко-русский, русско-немецкий, французско-русский, русско-французский, испанско-русский, русско-испанский, итальянско-русский, русско-итальянский, польско-русский, русско-польский, украинско-русский, русско-украинский, белорусско-русский, русско-белорусский и многоязычный параллельные корпуса;
* корпус диалектных текстов, включающий запись диалектной речи различных регионов России с сохранением их грамматической специфики; предусмотрен специальный поиск с учётом диалектной морфологии;
* корпус поэтических текстов, в котором возможен поиск не только по лексическим и грамматическим, но и по специфическим для стиха признакам (поиск определённого сочетания в сонетах, в эпиграммах, в стихотворениях, написанных амфибрахием, с определённым типом рифмовки и т. п.);
* обучающий корпус русского языка — корпус со снятой омонимией, разметка которого ориентирована на школьную программу русского языка;
* корпус устной речи, включающий расшифровки магнитофонных записей публичной и частной устной речи, а также транскрипты кинофильмов;
* акцентологический корпус (корпус истории русского ударения) — тексты, несущие информацию об истории русского ударения (все тексты поэтического корпуса и акцентуированные записи устной речи, в том числе кинофильмов). Эти тексты доступны для поиска по месту ударения и просодической структуре слова;
* мультимедийный корпус, куда входят снабжённые видео- и аудиорядом фрагменты кинофильмов 1930—2000-х годов. Возможен поиск не только по произносимому тексту, но и по жестам (кивание головой, похлопывание по плечу и т. п.) и типу речевого действия (согласие, ирония и т. п.)

# Литература

1. http://www.ruscorpora.ru/index.html
2. http://www.lomonosov-fund.ru
3. https://ru.wikipedia.org
4. http://www.lomonosov-fund.ru
5. https://myfilology.ru