Министерство образования и науки Российской Федерации

ФГБОУ ВО «Тверской государственный университет»

Факультет иностранных языков и международной коммуникации

Направление «Лингвистика»

Кафедра английского языка

РЕФЕРАТ

на тему: «Основные виды корпусов»

по дисциплине «Информационные технологии в лингвистике»

|  |  |
| --- | --- |
|  | Выполнила:  студентка 1 курса группы 12  дневной формы обучения  Маслобоева Татьяна Павловна |
|  |  |
|  | Проверила:  к.ф.н., доцент, кафедры английского языка  Масленникова Евгения Михайловна |

ТВЕРЬ

2021

**ВВЕДЕНИЕ**

**Корпусы и корпусная лингвистика**

***Корпусная лингвистика*** – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с использованием компьютерных технологий. Под названием ***лингвистический, или языковой, корпус текстов*** понимается большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач. В понятие «корпус текстов» входит также система управления текстовыми и лингвистическими данными, которую в последнее время чаще всего 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 результатов пользователю в удобной форме.

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.

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. Kucera) спроектировали его как набор из пятисот двухтысячесловных 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 последовали Ланкастерский корпус английского языка (Lancaster-Oslo-Bergen Corpus, LOB), Уппсальский корпус русского языка. Среди современных корпусов английского языка наиболее известны Британский национальный корпус (British National Corpus), Международный корпус английского языка (International Corpus of English), лингвистический Банк английского языка (Bank of English) и др. В настоящее время корпусы созданы для многих языков мира. Ведется работа и над созданием Национального корпуса русского языка.

В первой половине 90-х гг. корпусная лингвистика окончательно сформировалась как отдельный раздел науки о языке. При этом она тесно взаимодействует с компьютерной лингвистикой, используя ее достижения и в свою очередь обогащая ее.

Поиск в корпусе данных позволяет по любому слову построить конкорданс – список всех употреблений данного слова в контексте со ссылками на источник. 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 единиц и т.д. 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, и т.д. Корпусы призваны служить также источником и инструментом многоаспектных лексикографических работ по подготовке разнообразных исторических и современных словарей. Данные корпусов могут быть использованы для построения и уточнения грамматик и в целях обучения языку.

**Основные виды 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).

|  |  |
| --- | --- |
| **Признак** | **Тип данных** |
| **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 |
| **Объем текстов** | Полнотекстовые  «Фрагментнотекстовые» |
| **Хронологический аспект** | Синхронические  Диахронические |
| **«Общность»** | Общие  Одного писателя |
| **Структура** | Центральные и архивные  Ядерные и периферийные |

Таблица 1. Основные виды корпусов

Итак, *по типу языковых данных*корпусы делятся на письменные, устные и смешанные. В письменных корпусах устная речь не представлена (Брауновский корпус, LOB), в устных корпусах представлена только устная речь, смешанными обычно бывают национальные корпусы, представляющие бытование языка в определенный период времени (НКРЯ, BNC и др.).

*По критерию параллельности*корпусы делятся на одноязычные, двуязычные и многоязычные. В одноязычных корпусах противопоставляются диалекты, варианты языка. Например, такие разновидности английского языка, как английский как родной и английский как иностранный оставались за пределами научного интереса до появления новых технологий, позволивших вовлечь в контрастивный анализ существенно большее количество сопоставляемых произведений речи. Двуязычные и многоязычные корпусы объединяют тексты из одной и той же тематической области, независимо написанные на двух или нескольких языках (например, корпус материалов конференций по определенной научной проблеме, проходивших в разных странах и на разных языках). Такие корпусы помогают в работе с терминологией и часто используются переводчиками. Еще один вариант двуязычного или многоязычного корпуса – множество текстов-оригиналов, написанных на каком-либо исходном языке, и текстов-переводов этих исходных текстов на один или несколько других языков. Такой корпус предоставляет неоценимый материал для проведения сравнительносопоставительных исследований, для исследований по теории перевода и для обучения переводу человека и компьютера.

*По критерию «литературности» выделяются литературные,*

диалектные, разговорные, терминологические и смешанные корпусы. Примером разговорного корпуса может быть корпус Один Речевой День (ОРД), разрабатываемый в Санкт-Петербурге, примером терминологического корпуса – корпус текстов по корпусной лингвистике, позволяющий разрабатывать терминологический словарь непосредственно на живом текстовом материале. В этом корпусе методология корпусной лингвистики применена к ней самой.

*По цели*создания корпусы делятся на многоцелевые и специализированные. Многоцелевые корпусы обычно содержат тексты различных жанров (сюда относятся национальные корпусы), в то время как специализированные корпусы могут ограничиваться одним жанром или группой жанров.

Корпусы текстов могут быть классифицированы *по жанрам*и подразделяться на литературные, фольклорные, драматургические, публицистические и др. Примерами публицистического корпуса могут служить Компьютерный корпус текстов русских газет конца ХХ-ого века и корпус политических метафор.

Важным критерием для пользователей корпуса является его *доступность*. Свободно доступные корпусы позволяют в любое время в режиме on-line иметь доступ ко всем текстам корпуса в полном объеме. В ряде случаев свободный доступ может предоставляться к части корпусных данных. В работе с коммерческими корпусами нужно покупать право его использования on-line или копию на компакт-диске. Предварительно можно ознакомиться с аннотацией к корпусу или, возможно, даже поработать с корпусом в пробном режиме, но, как правило, не со всеми текстами, а только с небольшим по объему подкорпусом. Закрытые корпусы создаются для узко специфических целей и не предназначены для публичного использования.

*По назначению*выделяют исследовательские и иллюстративные корпусы. Исследовательские корпусы создаются с целью изучения различных аспектов функционирования языка. Этот тип корпусов ориентирован на широкий класс лингвистических задач. Неспецифицированность задачи требует при построении исследовательских корпусов использовать пропорциональное сужение, являющееся наиболее простым способом обеспечения репрезентативности. Как правило, такие корпусы текстов содержат от нескольких десятков миллионов до сотен миллионов словоупотреблений. Иллюстративные корпусы создаются после проведения научного исследования: их цель не столько выявить новые факты, сколько подтвердить и обосновать уже полученные результаты. Они служат для выделения из них лингвистических примеров, подтверждающих те или иные языковые (речевые, текстовые) факты, обнаруженные ранее иными лингвистическими приемами. Типичный пример иллюстративного корпуса представлен в «Путеводителе по дискурсивным словам русского языка», где семантический анализ частиц и выделенные значения сопровождаются значительным текстовым материалом, позволяющим читателю проверить предложенные семантические интерпретации.

Критерий *«динамичность»*подразделяет корпусы на динамические и статические. Первоначально корпусы текстов создавались как статические образования, отражающие определенное временнóе состояние языковой системы. Статические корпусы содержат тексты какого-то небольшого временнóго промежутка. Типичными представителями этого вида корпусов являются авторские корпусы – коллекции текстов писателей. Однако значительная часть чисто лингвистических и не только лингвистических задач требует выявления функционирования языковых феноменов на временнóй шкале – например, изменения значения слов, частоты использования тех или иных синтаксических конструкций и т.д. Для отражения процессуального аспекта проблемной области была разработана новая технология построения и эксплуатации динамического корпуса текстов. Динамические корпусы называют также мониторными или мониторинговыми. Цель мониторных корпусов – «складировать» постоянно растущее количество текстов в памяти компьютера. В течение заранее фиксированного промежутка времени происходит обновление и/или дополнение множества текстов корпуса. Неограниченные (постоянно развивающиеся) мониторные корпусы играют огромную роль в строении словаря, поскольку позволяют лексикографам следить за новыми словами, проникающими в язык, или за уже существующими словами, меняющими свое значение, а также за балансом их употребления в соответствии со стилем. В динамические корпусы текстов, как правило, включают письменные источники большого временнóго периода. Они предназначены для проведения различных диахронических исследований.

*Критерий «разметка»*делит корпусы на размеченные и неразмеченные. Существуют и другие термины, обозначающие это деление: индексированные и неиндексированные, аннотированные и неаннотированные, таггированые и нетаггированные. В размеченном корпусе словам или предложениям присваиваются метки (тэги) в соответствии с *характером разметки*: морфологические, синтаксические, семантические, просодические и др.

*По критерию «объем текстов»*выделяют полнотекстовые и так называемые фрагментотекстовые корпусы. Как известно, Брауновский корпус и корпус Ланкастер-Осло-Берген должны были строго соответствовать определенным критериям, одним из которых была длина текста, равная 2000 слов (словоупотреблений). Очевидно, что текстов, строго соответствующих таким критериям, практически нет. Следовательно, эти корпусы являются фрагментотекстовыми. К полнотекстовым корпусам относятся некоторые корпусы текстов определенного автора, а также корпусы коротких текстов, например, корпус мерфизмов (так называемых «законов подлости») или корпус газетных заголовков.

**Список используемой литературы**

Корпусы и корпусная лингвистика. Основные понятия. — Текст : электронный // Myfilology.ru – информационный филологический ресур – URL: https://myfilology.ru//177/korpusy-i-korpusnaya-lingvistika-osnovnye-ponyatiya/ (Дата обращения: 11.05.2021)

Захаров В.П, Богданова С.Ю. З-38 Корпусная лингвистика: учебник для студентов гуманитарных вузов. – Иркутск: ИГЛУ, 2011. – 161 с.