Министерство образования и науки Российской Федерации

ФГБОУ ВО «Тверской государственный университет»

Факультет иностранных языков и международный коммуникации

Направление «Лингвистика»

Кафедра английского языка

РЕФЕРАТ

на тему: «Типы корпусов текста»

по дисциплине «Информационные технологии в лингвистике»

|  |  |
| --- | --- |
|  | Выполнила: студентка 1 курса группы 12дневной формы обученияТашкевич Эллина Руслановна |
|  |  |
|  | Проверила:к.ф.н, доцент, кафедры английского языкаМасленникова Евгения Михайловна |

ТВЕРЬ

2021

**Содержание**

|  |  |
| --- | --- |
| Понятие корпусной лингвистики …………………………………..…… | 2 |
| Типы корпусов …………………………………………………………… | 3 |
| Специальные виды корпусов. Параллельный корпус. Сопоставительный корпус ………………………………………………………………… | 5 |
| Корпус текстов обучающихся …………………………………………… | 5 |
| Список используемой литературы ……………………………………… | 5 |

**Понятие корпусной лингвистики**

Корпусная лингвистика – раздел прикладной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов с применением 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.

Целесообразность создания и смысл использования корпусов определяется следующими предпосылками:

* достаточно большой (репрезентативный) 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, многими исследователями и в различных целях.

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.

Задача создателей корпуса – собрать как можно большее количество текстов, относящихся к тому подмножеству языка, для изучения которого корпус создается. Но главное не только и не столько в количестве языкового материала, сколько в его пропорциональности. Можно сказать, что корпус – это уменьшенная модель языка или подъязыка. Важнейшее понятие корпусной лингвистики – репрезентативность. Под репрезентативностью понимается необходимо-достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов и т.п.

**Типы корпусов**

Несмотря на разнообразие корпусов, можно выделить два основных способа деления корпусов на классы:

* 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.);
* разделение корпусов по типу лингвистической разметки. Несмотря на наличие множества типов разметки, большинство реально существующих корпусов относится к корпусам морфологического либо синтаксического типа. При этом следует подчеркнуть, что корпус с синтаксической разметкой явно или неявно включает в себя и морфологические характеристики лексических единиц.

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.

Прежде всего, корпусы различны по охвату материала. По этому признаку можно выделить:

* национальные корпусы, включающие тексты различных стилей и жанров (обычно в определенных временных рамках); это смешанные корпусы: они содержат образцы как литературной, так и разговорной речи;
* корпусы, представляющие какой-либо подъязык: только язык художественной литературы или, наоборот, диалектную речь, язык СМИ, язык определенной возрастной или социальной группы, язык одного писателя и т.п.; такие корпусы могут быть очень узкими по обхвату материала.

По полноте представления текстов корпусы делятся на:

* полнотекстовые (содержат полные тексты);
* фрагментированные (включают только фрагменты текстов определенной длины).

Можно классифицировать корпусы по типу лингвистической разметки. Прежде всего, корпусы могут быть не размеченными. Большинство существующих корпусов содержат только морфологическую разметку. Но есть и корпусы синтаксического типа (корпус с синтаксической разметкой не может не включать в себя и морфологические характеристики лексических единиц), а в ряде корпусов можно найти и семантическую, просодическую и другие типы разметок.

По форме хранения и представления данных можно выделить:

* письменные (только письменные тексты);
* мультимедийные (аудио- или видеозаписи);
* смешанные корпусы (аудио- или видеозаписи и письменные тексты).

По доступности корпусы делятся на:

* свободно доступные;
* коммерческие;
* закрытые.

По назначению:

* исследовательские корпусы, содержащие материал для лингвистических исследований (подавляющее большинство корпусов);
* иллюстративные корпусы, созданные с лингвометодическими целями – для обучения языку как иностранному.

По динамичности, скорости фиксации лингвистических процессов можно выделить:

* статичные корпусы; они не пополняются вообще (это референтные корпусы, в них включен весь существующий материал определенного рода) или постепенно пополняются различными материалами – не только новыми, но и созданными в предшествующие периоды;
* мониторинговые корпусы, которые позволяют следить за языком почти в режиме реального времени.

По хронологическому подходу корпусы делятся на:

* синхронические;
* диахронические.

По языку текстов корпусы разделяются на:

* одноязычные (русский, английский и т.д.);
* двуязычные (параллельные и сопоставительные);
* многоязычные (параллельные и сопоставительные).

**Специальные виды корпусов. Параллельный корпус. Сопоставительный корпус.**

Многоязычные корпусы могут состоять из разных текстов или из оригинальных текстов и их переводов на разные языки. Второй тип ресурсов называется параллельными корпусами. Существует также сопоставительный корпус, в котором собраны не переводные, а однотипные тексты на разных языках или вариантах языка. Это тексты на одну тему или одного жанра и т.д.

**Корпус текстов обучающихся.**

Корпус текстов обучающихся, или корпус не носителей языка – это электронная коллекция текстов, созданных не носителями данного языка, то есть людьми, изучающими иностранный язык. Корпус таких текстов называют корпусом ученических текстов, или просто ученическим корпусом.

**Список используемой литературы**

Информационный филологический ресурс (Myfilology.ru) – URL: <https://myfilology.ru> (дата обращения: 25.05.21)

Лекция 2. Типология корпусов и направления их использования: <https://docplayer.ru/88316315-Lekciya-2-tipologiya-korpusov-i-napravleniya-ih-ispolzovaniya.html> (дата обращения 25.05.21)