

МИНОБРНАУКИ РОССИИ
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО
ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
“ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ”

Лингвистические корпусы. Корпусная лингвистика

Подготовила: Усачева Валерия Сергеевна
студентка 1 курса, немецкого отделения, кафедры теории и методики
преподавания иностранных языков и культур

Определения и понятия

Что такое корпусная лингвистика?

Лингвистические корпусы - совокупность текстов, собранных в соответствии с определёнными принципами, размеченными по определённому стандарту и обеспеченные специализированной поисковой системой.

Корпусная лингвистика - раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с применением компьютерных технологий.

Цели и задачи

Цель курсовой работы: рассмотреть основные функции корпусной лингвистики и выявить практическую значимость данной области языкознания: возможности использования корпусной лингвистики в процессе изучения как русского, так и иностранного языков, а также в проведении лингвистических исследований.

Основные задачи курсовой работы:

1. Изучить определения основных понятий «корпуса языка» и корпусной лингвистики; выявить их типологию.
2. На основе исторических данных рассмотреть общие закономерности в становлении и развитии лингвистических корпусов.
4. Описать, каким образом можно использовать материалы Национального корпуса русского языка при изучении лексики, морфологии, синтаксиса.

Корпусная лингвистика как раздел прикладной ЛИНГВИСТИКИ

Одним из основных источников языкового материала является текст. Развитие вычислительной техники способствовало тому, что большое количество текстов стало доступно в электронном виде.

Задачи корпусной лингвистики

1. Создание и разметка корпусов текстов и разработка средств поиска по ним.
2. Экспериментальные исследования на базе корпусов.

Корпусная лингвистика как раздел прикладной ЛИНГВИСТИКИ

Корпус языка — это электронное собрание текстов, снабженное научным аппаратом.

Аппарат, «встроенный» в корпус, обычно называется «**разметкой**», или «аннотацией», корпуса; корпус тем лучше, чем полнее и совершеннее его аннотация.

Корпусная лингвистика как раздел прикладной ЛИНГВИСТИКИ

Разметка подразумевает приписывание текстам меток:

1. Экстралингвистические метки
 2. Лингвистические метки
- Морфологическая разметка
 - Синтаксическая разметка
 - Акцентная разметка
 - Семантическая разметка

Национальный корпус русского языка.

Использование корпуса на примере слова «школа»

Основной корпус [инструкция](#) [задать подкорпус](#) [English](#)

Поиск точных форм ?

Слово или фраза

Лексико-грамматический поиск ?

Слово ? <input type="button" value="А"/> <input type="button" value="Б"/> <input type="button" value="В"/>	Грамм. признаки ? выбрать	Семант. признаки ? выбрать
<input type="text"/>	<input type="text"/>	<input type="text"/>
Доп. признаки ? выбрать	Словообразование выбрать	<input checked="" type="checkbox"/> 1-е знач. <input checked="" type="checkbox"/> др. знач. <input type="checkbox"/> фильтр 1 <input type="checkbox"/> фильтр 2 ?
<input type="text"/>	<input type="text"/>	

Расстояние: от до ?

Слово ? <input type="button" value="А"/> <input type="button" value="Б"/> <input type="button" value="В"/>	Грамм. признаки ? выбрать	Семант. признаки ? выбрать
<input type="text"/>	<input type="text"/>	<input type="text"/>
Доп. признаки ? выбрать	Словообразование выбрать	<input checked="" type="checkbox"/> 1-е знач. <input checked="" type="checkbox"/> др. знач. <input type="checkbox"/> фильтр 1 <input type="checkbox"/> фильтр 2 ?
<input type="text"/>	<input type="text"/>	

Национальный корпус русского языка.

Использование корпуса на примере слова «школа»

Школа	
Лемма	школа (см. в словарях)
Грамматика	сущ, неод, ж, ед, им, disamb
Семантика основная	n:concr, t:org
Семантика дополнительная	n:abstr, t:fam
Доп. признаки	animred, capital, first, numred, posred
Сообщить об ошибке...	

Разметка лексемы «школа»

История зарождения Национальных лингвистических корпусов Международные лингвистические корпуса

**Брауновский лингвистический корпус,
1963**



**Британский национальный корпус,
1990**



История зарождения Национальных лингвистических корпусов.

Международные лингвистические корпуса



НАЦИОНАЛЬНЫЙ КОРПУС
РУССКОГО
ЯЗЫКА

Национальный корпус русского языка - крупнейшее электронное собрание текстов, включающее более 500 млн словоупотреблений. Создан в 2004 году.

История зарождения Национальных лингвистических корпусов

основной

— корпус

— биграммы

— триграммы

— 4-граммы

— 5-граммы

синтаксический

газетный

параллельный

обучающий

диалектный

поэтический

устный

акцентологический

мультимедийный

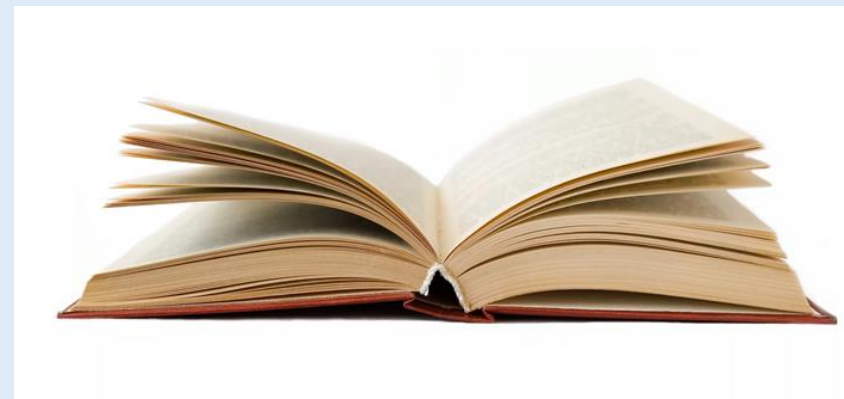
мультипарк

исторический

Корпусная лингвистика как способ изучения языка

Основные проблемы в развитии корпусной лингвистики

1. Объем материала.
2. Сложность в создании разметки.
3. Не выработаны единые подходы к классификации корпуса текстов. Такие классификации приводятся по различным основаниям зависимости от типа текстов, их языка, степени предварительной подготовки текстов.



Корпусная лингвистика как способ изучения языка

Основные перспективы в развитии корпусной лингвистики

Решение целого ряда лингвистических задач:


1. В лексикографии и лексикология (составление различных словарей, определение значений многоязычных слов)
2. В грамматике (определение частоты употребления грамматических морфем в текстах различного типа, определение частоты употребления классов слов и т.д.).

Корпусная лингвистика как способ изучения языка

Основные перспективы в развитии корпусной лингвистики

3. В лингвистике текста (дифференциация типов текста, выявление связи между предложениями в абзацах и между абзацами и т.д.).
4. При автоматическом переводе текстов (поиск контекстов слов, имеющих несколько переводных эквивалентов, фразеологических словосочетаний в параллельных текстах и т.д.).
5. В учебных целях (выбор цитат, отдельных произведений, примеров, используемых в процессе создания учебников и учебных пособий и т.д.).

Корпусная лингвистика как способ изучения языка

 НАЦИОНАЛЬНЫЙ КОРПУС
РУССКОГО
ЯЗЫКА

[перейти на страницу поиска](#) [выбрать подкорпус](#) [версия с ударениями](#) [настройки](#) [формат KWIC](#) [English](#)

Результаты поиска в **параллельном корпусе**

Объем всего корпуса: 1 953 документа, 5 844 884 предложения, 76 759 952 слова.

литература

Найдено 158 документов, 300 вхождений.

Поискать в других корпусах: [основном](#), [акцентологическом](#), [газетном](#), [диалектном](#), [мультимедийном](#), [обучающем](#), [поэтическом](#), [синтаксическом](#), [устном](#).

Страницы: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [11](#) [следующая страница](#)

1. César Cervera. «Mussolini jamás acabó con la Mafia. Eso fue un mito de la propaganda fascista» [www.abc.es] (2015.10.26) [омонимия не снята] [Все примеры \(1\)](#)

es	—En relación a EE.UU, su cine y su literatura contribuyó a mitificar a la Mafia. [César Cervera. «Mussolini jamás acabó con la Mafia. Eso fue un mito de la propaganda fascista» [www.abc.es] (2015.10.26)] [омонимия не снята] ←...→
ru	— Если говорить о США, то их кинематограф и литература в значительной степени способствовали созданию мифов о мафии. [Муссолини вовсе не покончил с мафией. Это был лишь миф фашистской пропаганды (http://inosmi.ru/world/20151029/231074173.html, 29.10.2015)] [омонимия не снята] ←...→

Исследование лексемы «литература» в параллельных текстах

Вывод

С появлением корпусной лингвистики исчезли ограничения на объем анализируемого материала.

В распоряжении исследователя оказываются колоссальные массивы текстов самого разного типа.



МИНОБРНАУКИ РОССИИ
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО
ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
“ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ”

Лингвистические корпусы. Корпусная лингвистика

Подготовила: Усачева Валерия Сергеевна
студентка 1 курса, немецкого отделения, кафедры теории и методики
преподавания иностранных языков и культур