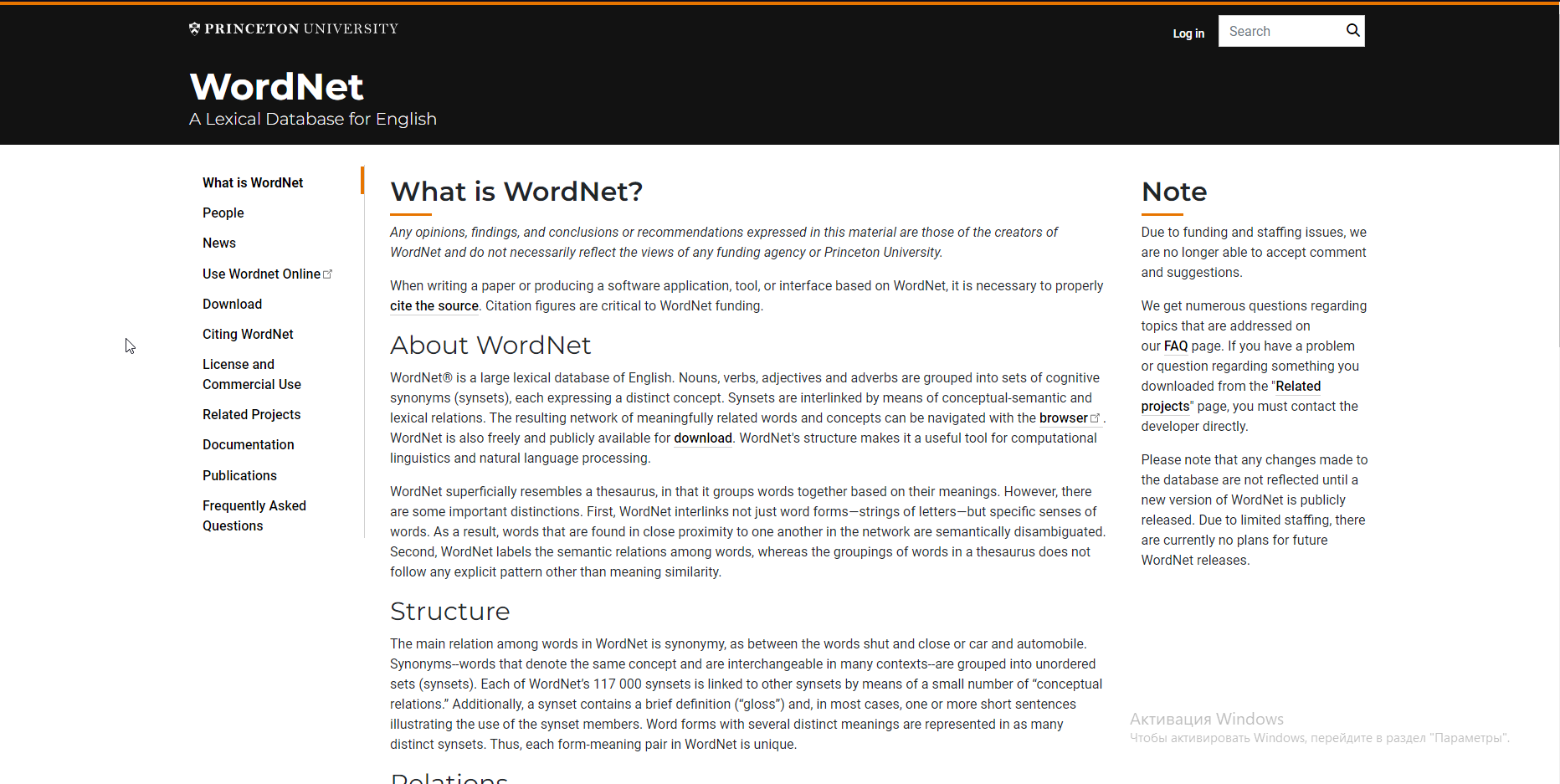
**Электронные лингвистические ресурсы.**

1. WordNet - <https://wordnet.princeton.edu>



Возможности:

* Определение синонимов и антонимов слов.
* Предоставление информации о связях между словами, таких как гиперонимия (отношение "выше-ниже") и гипонимия (отношение "ниже-выше").
* Классификация слов по семантическим категориям и группам.
* Предоставление примеров употребления слов в контексте.
* Поиск слов по их связям и значению.
* Построение онтологий и семантических сетей на основе лексической информации.
* Поддержка различных языков и семантических отношений между словами.

1. FrameNet - <https://framenet.icsi.berkeley.edu/fndrupal/>



Возможности:

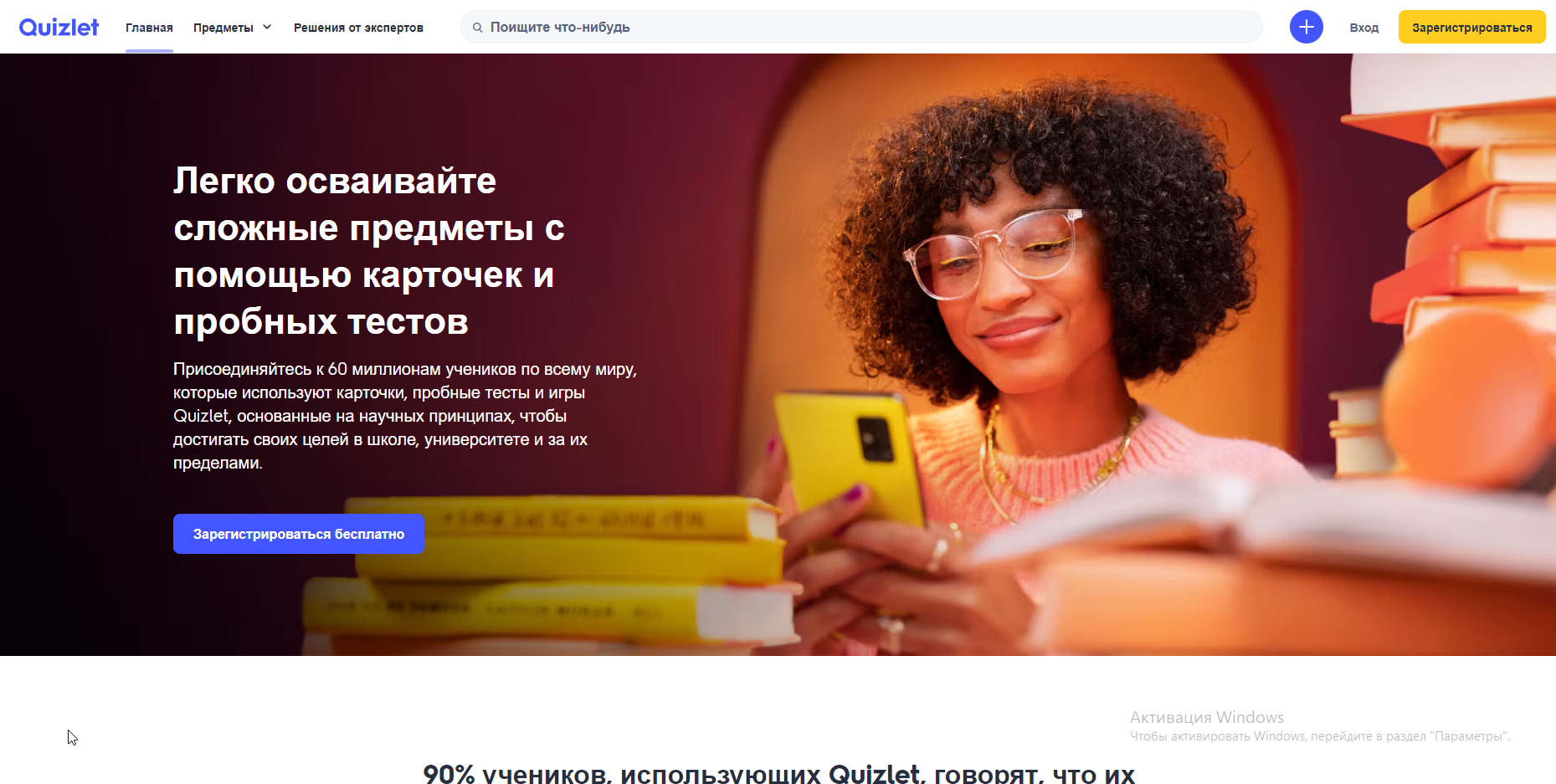
* Содержит информацию о семантических кадрах и лексических единицах английского языка.
* Позволяет исследователям изучать семантические отношения и значения слов.
* Предоставляет данные о том, как слова используются в контексте.
* Служит инструментом для анализа и понимания семантики языка.

1. eLibrary - <https://elibrary.ru/defaultx.asp>



Возможности: eLibrary предоставляет доступ к электронным книгам, статьям из научных журналов, аудиокнигам, видеоурокам, базам данных, каталогам библиотек, исследованиям и другим образовательным материалам.

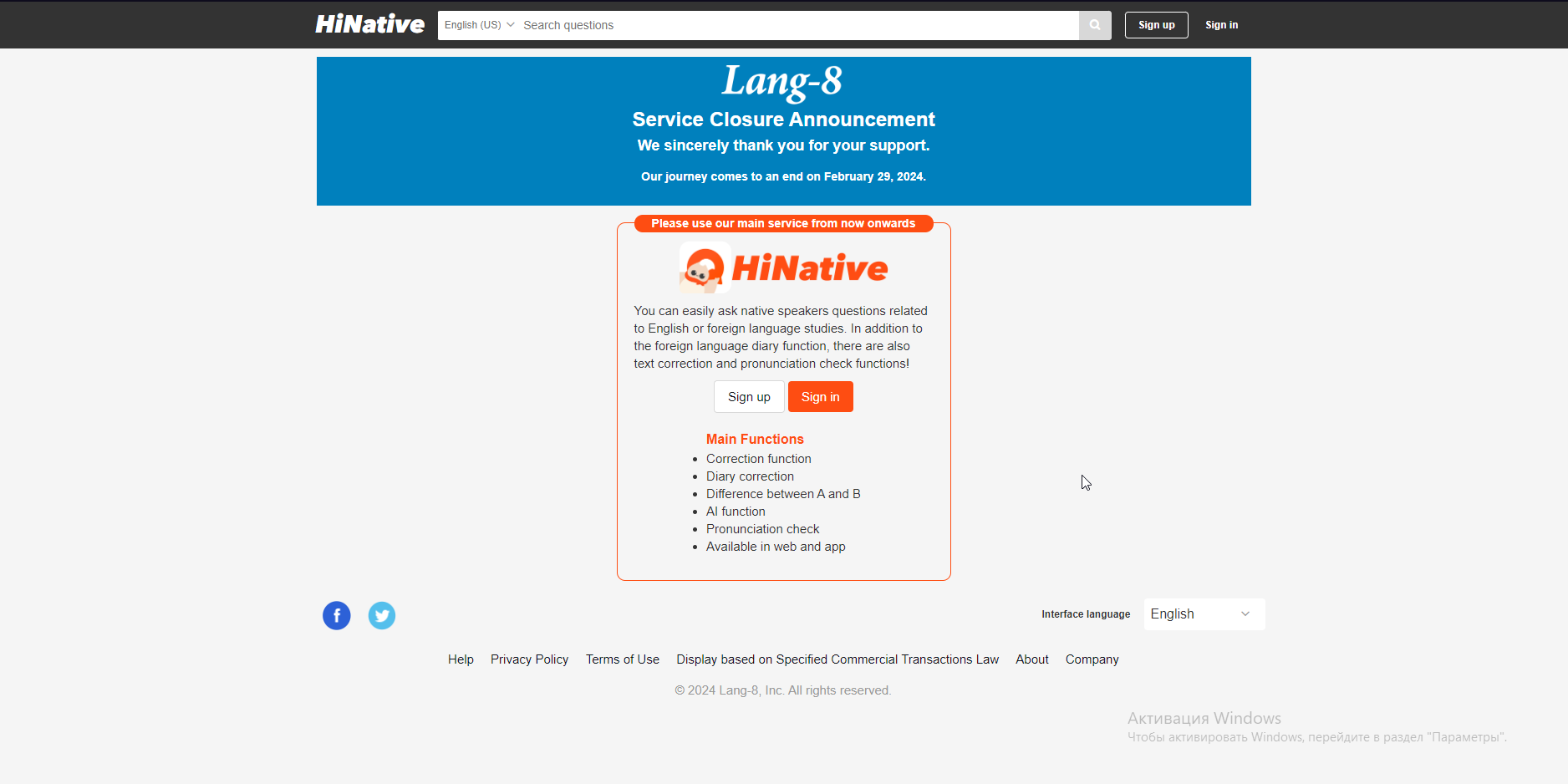
1. Quizlet - <https://quizlet.com>



Возможности:

* Тестировать память с функцией Заучивание
* Ставить рекорды в игре Подбор
* Делиться учебными модулями
* Слушать произношение слов на 18 языках
* Улучшать обучение с помощью изображений и аудио.

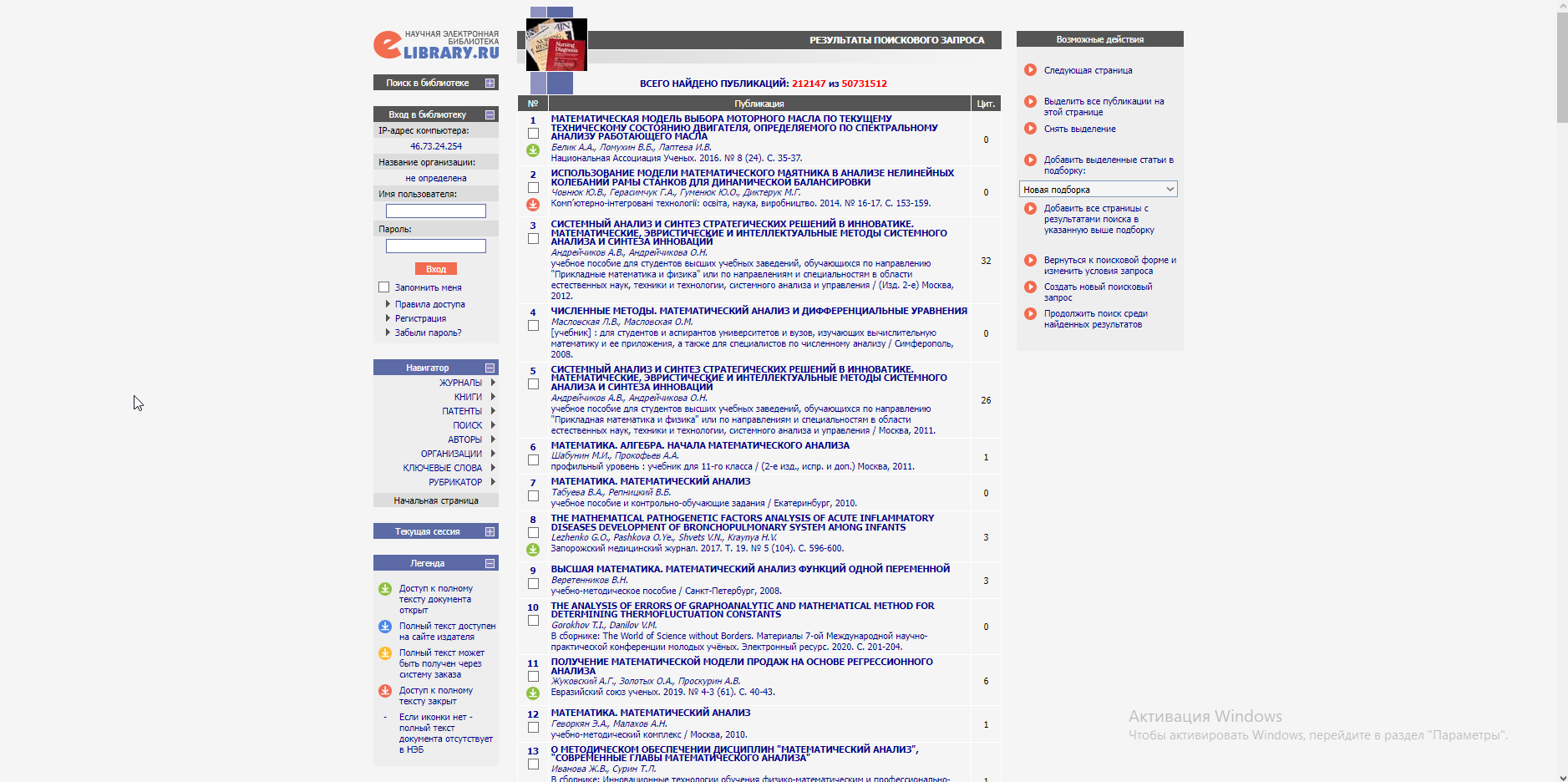
1. Lang – 8 - <https://hinative.com/lang-8>



Возможности: Lang-8 позволяет пользователям учить языки, публикуя тексты на выбранном языке и получая коррекцию от носителей языка. Можно писать тексты, исправлять других пользователей, общаться в комментариях, устанавливать цели по изучению языков, использовать ресурсы для улучшения навыков.

Я решил воспользоваться eLibrary. Выполнил следующий поисковой запрос: «Математический анализ».

Результат:



Вывод: Данная библиотека содержит огромное множество статей на всевозможные темы, при этом если более детально описать запрос, то можно найти нужную вам информацию намного быстрее. Такой ресурс очень полезен, так как такое количество различной информации мало где можно встретить в наше время, поэтому если вы хотите прочесть какой-либо источник, то стоит обратиться именно к этому ресурсу, потому что почти гарантированно вы найдете верную информацию.

1. Что может являться единицей корпуса? - Текстовый документ, предложение, абзац или другой текстовый блок.
2. Как отбираются тексты для корпуса? Проиллюстрируйте принципы отбора на примере Брауновского и других корпусов. - Тексты для корпуса отбираются на основе различных критериев, таких как репрезентативность языка, баланс жанров и тематик, общественная значимость и др. Примеры: Брауновский корпус создан на основе различных текстов американского английского языка; British National Corpus содержит тексты из Великобритании, жанрово и тематически разнообразных.
3. Представьте классификацию корпусов. Поясните, что означает «исследовательский корпус», «статический корпус», «параллельный корпус». - *Исследовательский корпус*: Корпус текстов, созданный специально для проведения научных исследований и анализа языка. *Статический корпус*: Коллекция текстов, представляющая собой фиксированный набор данных, обычно используемый для обучения и тестирования языковых моделей. *Параллельный корпус*: Корпус, содержащий тексты на двух или более языках, предназначенный для исследования и разработки машинного перевода и других языковых задач.
4. Охарактеризуйте Британский национальный корпус по следующим критериям: количество словоупотреблений, вид корпуса (по разным признакам). - Британский национальный корпус (BNC) содержит около 100 миллионов словоупотреблений и представляет собой корпус текстов, собранных из различных источников, таких как странные разговоры, литература, газеты и т. д. Этот корпус является репрезентативным с точки зрения широкого спектра жанров и стилей текстов, что делает его ценным для лингвистических исследований.
5. Составьте глоссарий по теме «Корпусная лингвистика». Используйте для этого рекомендуемые источники литературы и сетевые ресурсы. Включите в глоссарий определения следующих понятий: конкорданс, рандомизация, коллокация, подмассив, парсинг, лемматизация, корпус-менеджер. –

1. \*\*Конкорданс\*\*: Список конкретных слов или фраз из текста корпуса, с их контекстами для анализа.

2. \*\*Рандомизация\*\*: Процесс перестановки данных в корпусе для исключения влияния порядка на результаты анализа.

3. \*\*Коллокация\*\*: Слова или фразы, склонные к частому соседству друг с другом в корпусе, выражающие определенное значения или функциональные отношения.

4. \*\*Подмассив\*\*: Набор данных, полученных из корпуса по заданным параметрам или фильтрам для последующего анализа.

5. \*\*Парсинг\*\*: Анализ предложения на естественном языке для определения синтаксической структуры и семантики.

6. \*\*Лемматизация\*\*: Процесс приведения слова к его базовой форме (лемме) для унификации разных форм слова.

7. \*\*Корпус-менеджер\*\*: Программное обеспечение для работы с лингвистическими корпусами, включающее функции по поиску, анализу и управлению данными корпуса.

Источники:

- Manning, Christopher D., and Hinrich Schütze. "Foundations of statistical natural language processing." MIT press, 1999.

- McEnery, Tony, and Andrew Hardie. "Corpus linguistics: Method, theory and practice." Cambridge University Press, 2012.

6. Найдите сетевые ресурсы по теме «корпусная лингвистика» и кратко охарактеризуйте их. - Universal Dependencies (universaldependencies.org): Коллекция аннотированных лингвистических корпусов на различных языках, основанных на синтаксической разметке.

Corpus of Contemporary American English (corpus.byu.edu/coca): Крупный корпус современного американского английского языка, включающий тексты из различных источников, таких как новости, литература, разговорный язык.

Natural Language Toolkit (NLTK) (www.nltk.org): Библиотека для обработки естественного языка в Python, предоставляющая инструменты для работы с текстом, корпусами и лексиконами.

**Лабораторная работа 10.**

Задание 1.

|  |  |  |  |
| --- | --- | --- | --- |
|  | НКРЯ | COCA | BNC |
| Количество вхождений | В период с 2016 по 2021 год количество запросов изменялось в промежутке с 66 по 151 запрос. Самое минимальное количество запросов было сделано в 2021 году. | 200 | 169 |

По результатам количесвам вхождений мы выяснили, что больше всего вхождений на сайте COCA, а меньше всего у НКРЯ.

Задание 2.

|  |  |  |  |
| --- | --- | --- | --- |
| № примера | НКРЯ | COCA | BNC |
| 1 | Загадочная русская женщина(**Наши дети: Подростки (2004)**) | Андрей Ошарович Протопопов, Церковно-государственные отношения в **Российской** Федерации, 1996, BYU L. Rev. 853 (1996). Доступно по адресу: https://digitalcommons.law.byu ... | ... которые были заключены в квадратные скобки (например, [T] he...). Отчет Мюллера, официально озаглавленный Отчет о расследовании российского вмешательства…  https://www.sketchengine.eu/mueller-report-corpus/?highlight=Russian%20eat |
| 2 | Мое сознание постепеннно забыло что русский — это мой язык " by default "  [Запись LiveJournal (2004)] | [Эволюция **русской** языковой картины мира и **культура** речи](https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=1214&context=rlj) scholarsarchive.byu.edu " cgi " viewcontent | Корпус римской драмы (36 пьес) Корпус русской драмы…  https://www.sketchengine.eu/dracor-drama-corpora/?highlight=Russian |
| 3 | Оба они являются первыми переводами на русский язык текстов двух важнейших религиозно-философских систем Индии — санкхьи и йоги.  [В. Г. Лысенко. Классическая индийская философия в переводах и исследованиях российских ученых (1990-1996) (1996)] | Первый курс **русского** языка, часть 1... В этом введении в **русский** язык, культуру и историю студенты разовьют базовые навыки говорения, аудирования, чтения, ...  is.byu.edu " каталог " RUSS-041-T002 | ...Русские примеры использования, словосочетания-ов в русском языке, а также тезаурус, в котором есть синонимы и похожие слова. SKELL основан на…  https://www.sketchengine.eu/ruskell-examples-and-collocations-for-learners-of-russian/?highlight=Russian%20eat |

Задание 3.

|  |  |  |  |
| --- | --- | --- | --- |
| № примера | НКРЯ | COCA | BNC |
| 1 | прилагательное | прилагательное | прилагательное |
| 2 | существительное | прилагательное | прилагательное |
| 3 | прилагательное | прилагательное | прилагательное |

Задание 4.

|  |  |  |  |
| --- | --- | --- | --- |
| № примера | НКРЯ | COCA | BNC |
| 1 | Женщина русской национальности | Относится к политики РФ | Относится к политике России |
| 2 | Относится к русской нации | Относится к русской культуре | Относится к русской культуре |
| 3 | Относящийся к русскому языку | Относится к русскому языку | Относится к русскому языку |

Задание 5.

При сравнении морфологической и лексической характеристики слова "русский" в разных корпусах с разным контекстом, можно прийти к следующим выводам:

1. Морфологические характеристики: в разных контекстах слово "русский" может принимать разные формы в зависимости от своей роли в предложении. Например, в одном контексте это может быть форма прилагательного (русский язык), а в другом - форма существительного (русские люди).

2. Лексические характеристики: в разных контекстах слово "русский" может иметь разные значения или оттенки значений. Например, в одном контексте это может означать принадлежность к нации (русский человек), а в другом - относящийся к России (русский язык).

Таким образом, сравнение морфологической и лексической характеристики слова "русский" в разных контекстах позволяет увидеть его разнообразие использования и значение в зависимости от контекста.

Задание 6.

Корпусы, такие как НКРЯ (Национальный корпус русского языка), COCA (Corpus of Contemporary American English) и BNC (British National Corpus), являются крупными коллекциями текстов, используемыми для лингвистических исследований.

Использование этих корпусов позволяет исследователям проводить сравнительные анализы между различными языками, диалектами, жанрами текстов и другими лингвистическими аспектами. Такие корпусы представляют собой ценный инструмент для изучения языка и его использования в различных контекстах.

ы

1. В чём суть эмпирической методологии в лингвистическом анализе? - Эмпирическая методология в лингвистическом анализе основана на сборе и анализе конкретных языковых данных для вывода общих закономерностей в языке.
2. Дайте определение понятию коллокация (collocation). - Коллокация – это тесное сочетание двух или более слов, которое часто встречается вместе и образует естественную лексическую единицу.
3. Сформулируйте закон Ципфа. - Закон Ципфа утверждает, что если слова в тексте упорядочены по убыванию частоты их употребления, то ранг слова обратно пропорционален его частоте встречаемости.
4. Дайте определение понятию дискурсивной просодии (discourse prosody). - Дискурсивная просодия – это изучение интонации, ритма и других звуковых характеристик речи в контексте общения и восприятия информации.
5. Дайте определение понятию семантической просодии (semantic prosody). - Семантическая просодия – это эффект, когда слово или выражение несет скрытую эмоциональную или оценочную окраску, влияющую на восприятие и ассоциации.
6. В чём сущность гипотезы общего ядра (common core hypothesis). - Гипотеза общего ядра предполагает, что различные языки могут иметь общее ядро базовых концепций и категорий, которые являются универсальными для всех языков.
7. Что такое дерево зависимостей? Постройте дерево зависимостей для простого предложения английского языка. Постройте дерево зависимостей для сложного предложения английского языка. - Дерево зависимостей – это графическое отображение синтаксических связей между словами в предложении. В простом предложении: "The cat sleeps," дерево зависимостей будет иметь вид подобный:

sleeps (ROOT)

/ \

The (det) cat (nsubj)

Для сложного предложения: "I like to read books and watch movies," дерево зависимостей будет иметь вид подобный:

like (ROOT)

/ |

I read (xcomp)

/ \

to (aux) books (dobj)

like (ROOT)

/ \

I (nsubj) watch (conj)

/ \

to (aux) movies (dobj)

1. Назовите основные сферы применения корпусов? - Лингвистика, обработка естественного языка, машинное обучение, исследования в области компьютерной лингвистики.
2. Дайте синонимы термину аннотация (annotation). - обзор, резюме, анализ, комментарий.
3. Дайте определение понятию репрезентативность корпуса. - Репрезентативность корпуса – это степень, в которой корпус текста отражает разнообразие языка или языковые особенности определенной группы говорящих или текстов.
4. Дайте понятие определению кластерный анализ (cluster analysis). - Кластерный анализ – это метод статистического анализа, который используется для разделения набора данных на группы (кластеры) таким образом, чтобы объекты внутри одной группы были более похожи между собой, чем на объекты из других групп.
5. Что такое конкорданс? - Конкорданс – это упорядоченный список слов или фраз из текста с указанием контекста, который используется для анализа языка и выявления закономерностей в использовании слов.
6. Что такое конкордансер? - Конкордансер – это инструмент или программное обеспечение, которое позволяет искать и анализировать употребление слов или выражений в тексте, обеспечивая контекст и частоту их встречаемости.
7. Дайте определение понятию дистрибуция. Покажите дистрибуцию двух синонимов (на выбор). - Дистрибуция – это процесс распределения продукции от производителя к конечному потребителю через различные каналы сбыта.

Пример дистрибуции синонимов:

Распределение:

- Производитель -> Оптовик -> Розничный магазин -> Покупатель.

Сбыт:

- Производитель -> Дистрибьютор -> Торговая точка -> Конечный потребитель.

1. Дайте определение понятию частотность. Определите частотность любой лексической единицы английского языка по различным корпусам. - Частотность – это количество раз, которое определенное явление встречается в данном контексте или наборе данных. Например, в лингвистике, частотность относится к количеству вхождений конкретного слова или фразы в текст.

Для определения частотности лексической единицы английского языка можно использовать различные корпуса, такие как Google Ngram Viewer, Corpus of Contemporary American English (COCA), British National Corpus (BNC) или другие текстовые базы данных, которые позволяют анализировать употребление слов в реальных текстах и определять их частотность.

1. Дайте определение понятию лемма, лемматизация. - Лемма – это базовая форма слова. Лемматизация – это процесс приведения слова к его лемме.
2. Дайте определение понятию лексическая плотность. - Лексическая плотность – это соотношение числа уникальных слов к общему числу слов в тексте или речи, отражающее разнообразие и насыщенность лексикой.
3. Что такое морфологический анализатор. - Морфологический анализатор – это инструмент или программа, которая анализирует структуру слова, определяя его морфологические характеристики, такие как часть речи, падеж, число, время и т. д.
4. Дайте определение понятию сегментация корпуса. - Сегментация корпуса - это процесс разделения текста на более мелкие части, такие как предложения, абзацы или слова, для последующего анализа или обработки.