**Уровень 1**

**Задание 1.**

Для выполнения задания я взяла данные ресурсы:

- **The LINGUIST list**The LINGUIST List - это онлайн-ресурс, предоставляющий информацию о лингвистике и языках. На этом сайте можно найти объявления о конференциях, стажировках, академических вакансиях, публикациях, а также другие материалы и ресурсы, связанные с лингвистикой.

Для чего можно использовать ресурсы The LINGUIST List:

1. Поиск актуальной информации о событиях в мире лингвистики.

2. Поиск вакансий и возможностей для стажировок в области лингвистики.

3. Получение доступа к академическим публикациям и ресурсам по лингвистике.

4. Общение и обмен опытом с другими лингвистами.

Для производства запросов на сайте The LINGUIST List можно использовать поиск по ключевым словам, фильтры по категориям (например, конференции, вакансии, публикации) и другие опции, предоставляемые на сайте.

Синтаксис запроса на The LINGUIST List зависит от того, какую информацию вы ищете. Обычно можно использовать ключевые слова или фразы для поиска конкретной информации.

Преимущества The LINGUIST List:

1. Большое количество актуальной информации о лингвистике.

2. Возможность быстро найти интересующие материалы и ресурсы.

3. Возможность участвовать в академическом сообществе лингвистов.

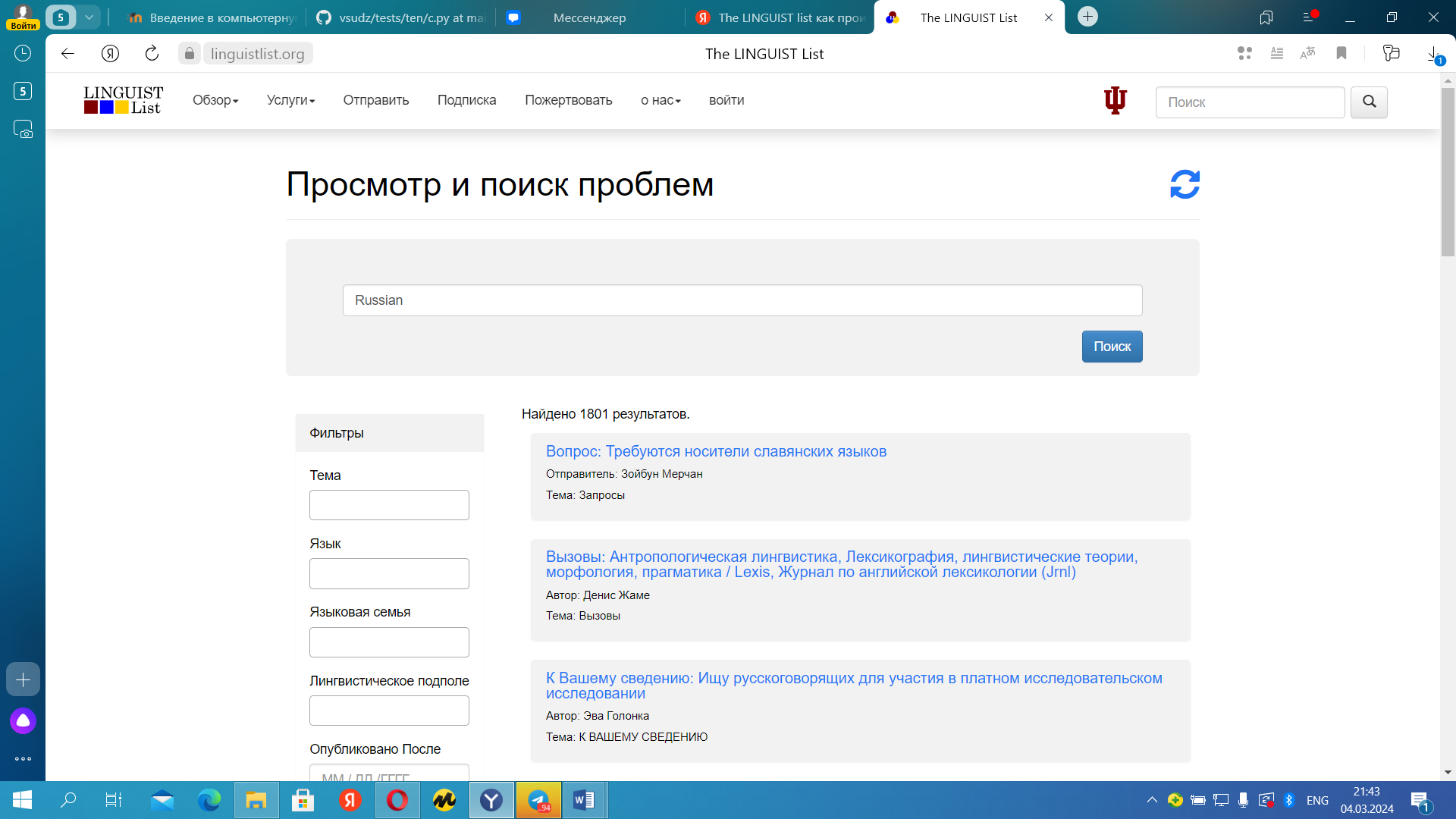
Недостатки The LINGUIST List:

1. Возможно, не всегда полная актуализация информации.

2. Некоторые ресурсы могут быть доступны только для платных подписчиков.

3. Не всегда удобный интерфейс для поиска и навигации.

Похожие ресурсы на The LINGUIST List могут включать другие онлайн-платформы для лингвистов, такие как Linguistic Society of America (LSA), Association for Computational Linguistics (ACL), или сайты университетов и исследовательских центров, специализирующихся на лингвистике.



Для проверки программы я ввела в поисковик запрос язык, про который хотела узнать. Впринципи сайт выдал статьи про указанный язык, поэтому можно сказать что сайт работает неплохо.

- **LingBuzz**LingBuzz - это онлайн-репозиторий для публикации и обмена научными статьями и документами в области лингвистики. На этом ресурсе представлена широкая гамма статей и работ по различным темам в области языкознания.

Для чего можно использовать ресурсы LingBuzz:

1. Поиск и доступ к актуальным научным статьям и документам по лингвистике.

2. Публикация собственных исследований и получение обратной связи от сообщества лингвистов.

3. Обмен опытом и знаниями с коллегами в области языкознания.

Для производства запросов на сайте LingBuzz можно использовать поиск по ключевым словам, фильтры по авторам, темам, датам публикации и другие опции, предоставляемые на сайте.

Синтаксис запроса на LingBuzz зависит от того, какую информацию вы ищете. Обычно можно использовать ключевые слова или фразы для поиска конкретной статьи или документа.

Преимущества LingBuzz:

1. Большой выбор научных статей и документов по лингвистике.

2. Возможность быстро найти интересующие материалы и ресурсы.

3. Возможность публикации собственных исследований и получения обратной связи от сообщества.

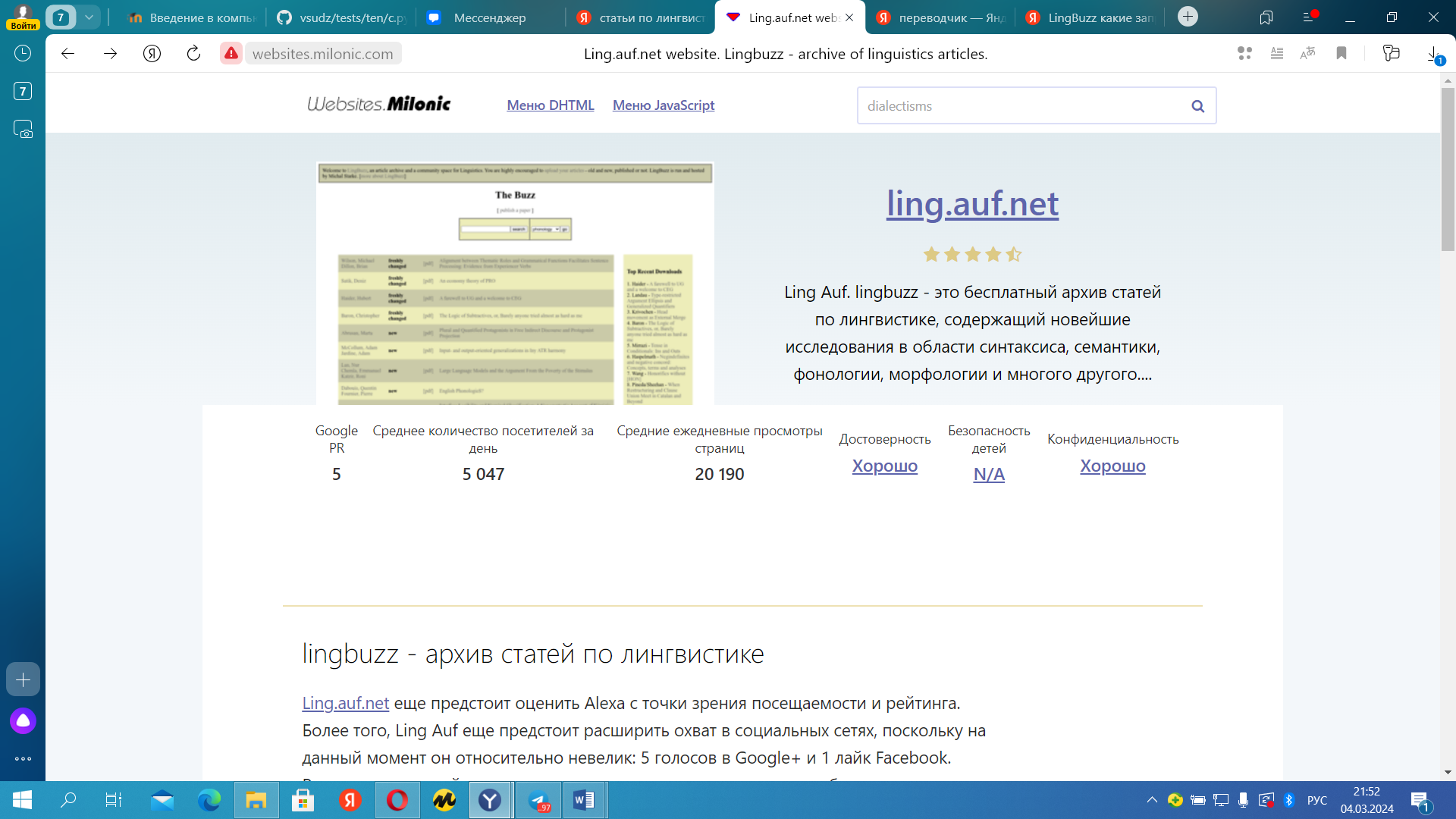
Недостатки LingBuzz:

1. Возможно, не всегда полная актуализация информации.

2. Некоторые материалы могут быть доступны только для зарегистрированных пользователей.

3. Не всегда удобный интерфейс для поиска и навигации.

Похожие ресурсы на LingBuzz могут включать другие онлайн-репозитории для публикации научных статей, такие как ResearchGate, Academia.edu, или сайты журналов по лингвистике.



Вбивая запрос на этот сайт в области языкознания, он выдаёт статьи по заданной теме. Сайт работает достаточно хорошо, но не все запросы сайт распознает. Например, при вводе запроса диалектизмы сайт ничего не выдаёт. А при вводе запроса науки о языке сайт предоставляет статьи по этой теме.

- eLIBRARY.RU

eLIBRARY.RU - это онлайн-библиотечная платформа, предоставляющая доступ к научным статьям, книгам, журналам и другим академическим материалам в различных областях знаний, включая лингвистику. На этом ресурсе представлена широкая коллекция публикаций от российских и зарубежных издательств.

Для чего можно использовать ресурсы eLIBRARY.RU:

1. Поиск и доступ к актуальным научным статьям, книгам и журналам по различным тематикам, включая лингвистику.

2. Использование материалов для научных исследований, учебы или профессионального развития.

3. Обновление информации и получение доступа к новым публикациям и исследованиям.

Для производства запросов на сайте eLIBRARY.RU можно использовать поиск по ключевым словам, фильтры по авторам, году публикации, типу материала и другие опции, предоставляемые на сайте.

Синтаксис запроса на eLIBRARY.RU может включать ключевые слова, фразы в кавычках для поиска точного выражения, операторы логических связей (AND, OR, NOT) для уточнения запроса.

Преимущества eLIBRARY.RU:

1. Большой выбор научных статей, книг и журналов по различным областям знаний.

2. Возможность быстрого доступа к актуальной информации и ресурсам.

3. Возможность сохранения понравившихся материалов и создания персональной библиотеки.

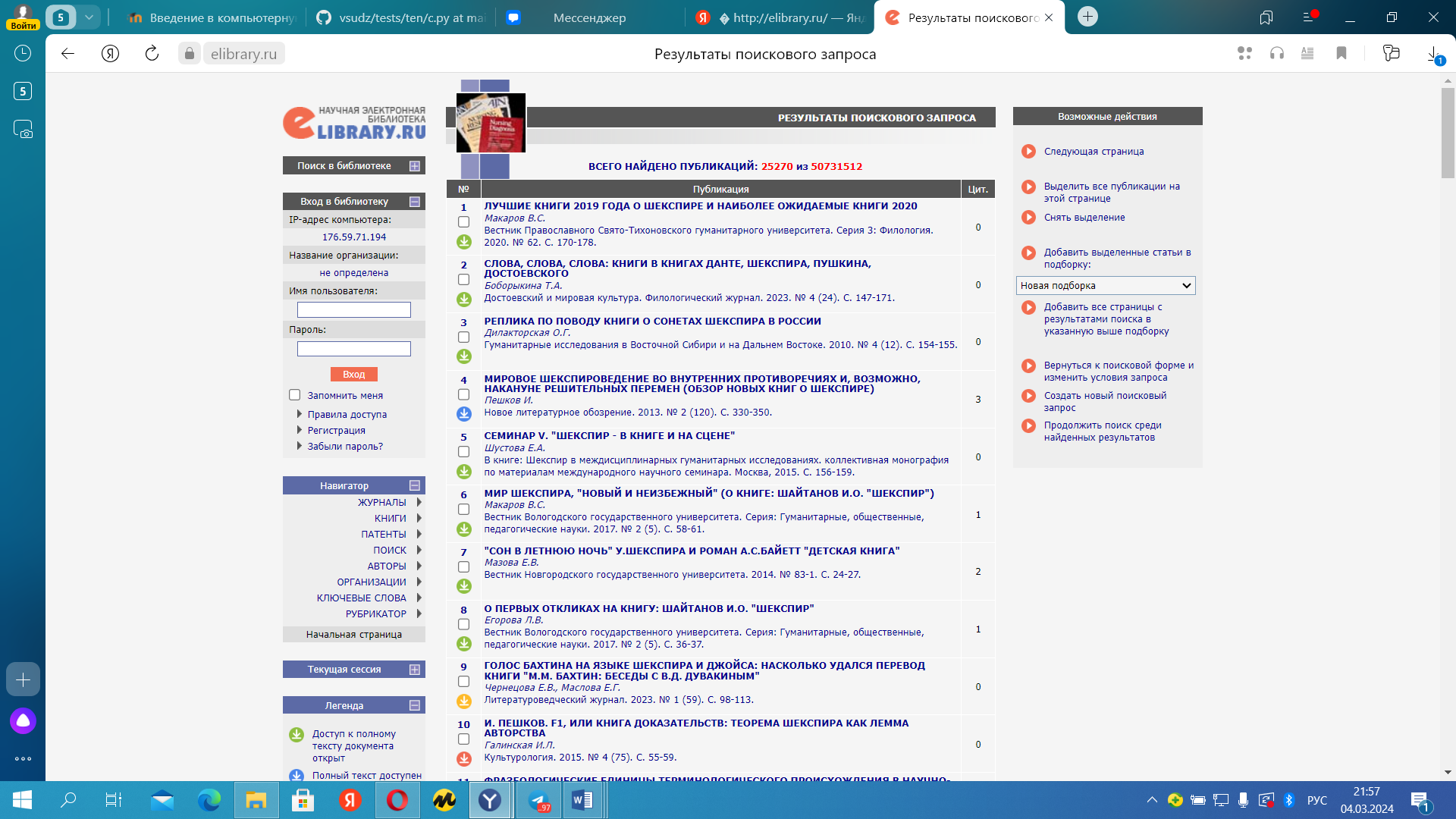
Недостатки eLIBRARY.RU:

1. Некоторые материалы могут быть доступны только по подписке или платно.

2. Не всегда полная актуализация информации или ограниченный доступ к некоторым материалам.

3. Не всегда удобный интерфейс для поиска и навигации по платформе.

Похожие ресурсы на eLIBRARY.RU могут включать другие онлайн-библиотечные платформы, такие как JSTOR, ScienceDirect, SpringerLink и др., предоставляющие доступ к научным публикациям в различных областях знаний.



Для проверки данного сайта я ввела запрос «книги Шекспира». И действительно сайт выдаёт его книги. Сайт работает хорошо, из всех выбранных сайтов этот мне понравился больше всего.

- Информационный блог об изучении иностранных языков.

Информационный блог об изучении иностранных языков - это онлайн-ресурс, на котором представлена информация о методиках изучения иностранных языков, полезные советы, рекомендации, обзоры учебных материалов, интересные факты о языках и культуре стран, а также другие темы, связанные с изучением иностранных языков.

Для чего можно использовать информационный блог об изучении иностранных языков:

1. Получение полезной информации о методах и стратегиях изучения иностранных языков.

2. Обмен опытом и советами с другими изучающими языки.

3. Ознакомление с новыми учебными материалами, онлайн-курсами и ресурсами для изучения языков.

4. Поддержка мотивации и вдохновение для продолжения изучения иностранных языков.

Для производства запросов на информационном блоге об изучении иностранных языков можно использовать поиск по ключевым словам или темам, а также навигационные разделы или категории на сайте.

Синтаксис запроса на информационном блоге может включать ключевые слова, фразы в кавычках для поиска точного выражения, операторы логических связей (AND, OR, NOT) для уточнения запроса.

Преимущества информационного блога об изучении иностранных языков:

1. Полезная и актуальная информация о методиках и ресурсах для изучения языков.

2. Возможность общения с единомышленниками и получение поддержки в процессе изучения.

3. Вдохновляющие и мотивирующие материалы для продолжения обучения.

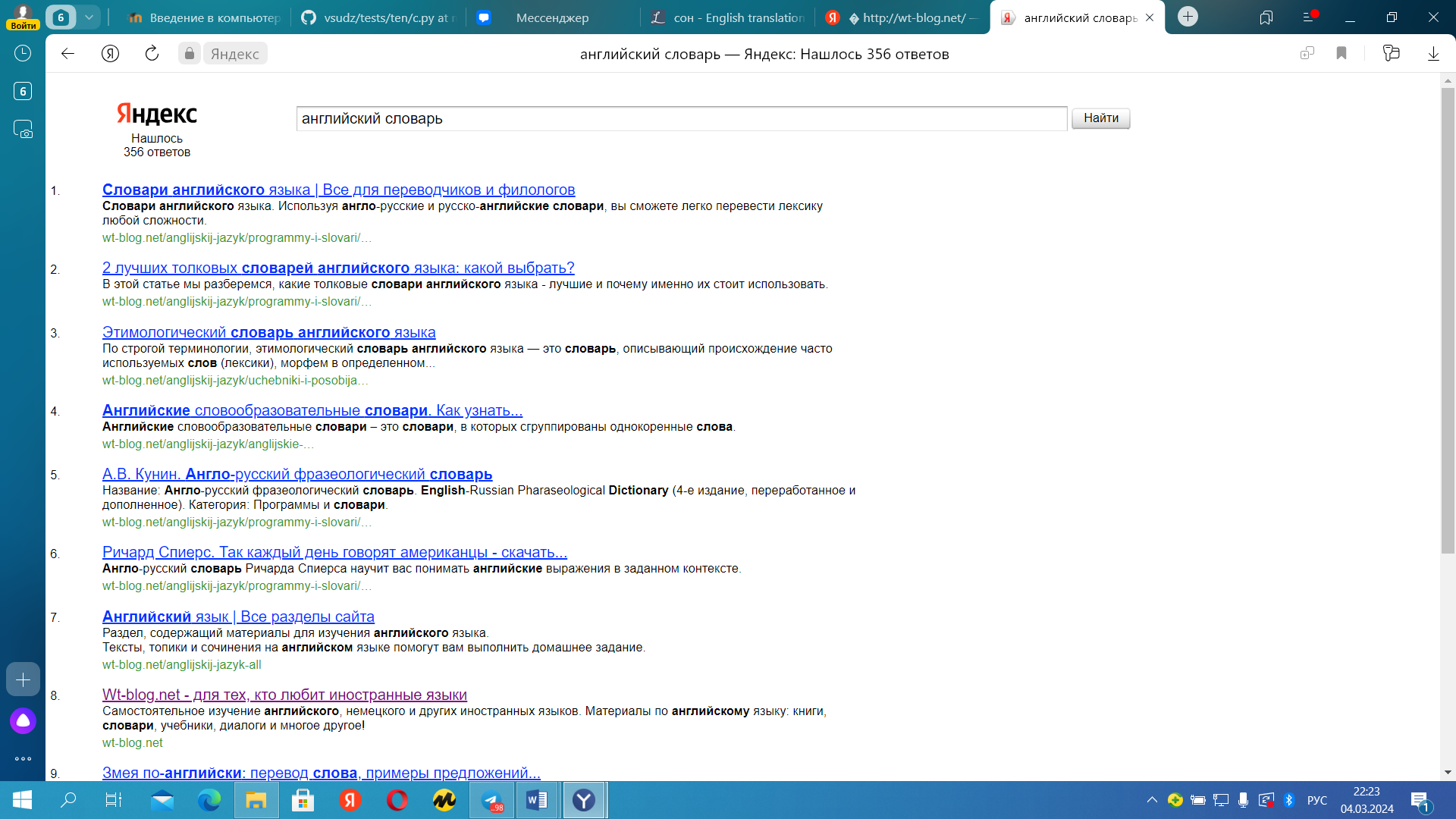
Недостатки информационного блога об изучении иностранных языков:

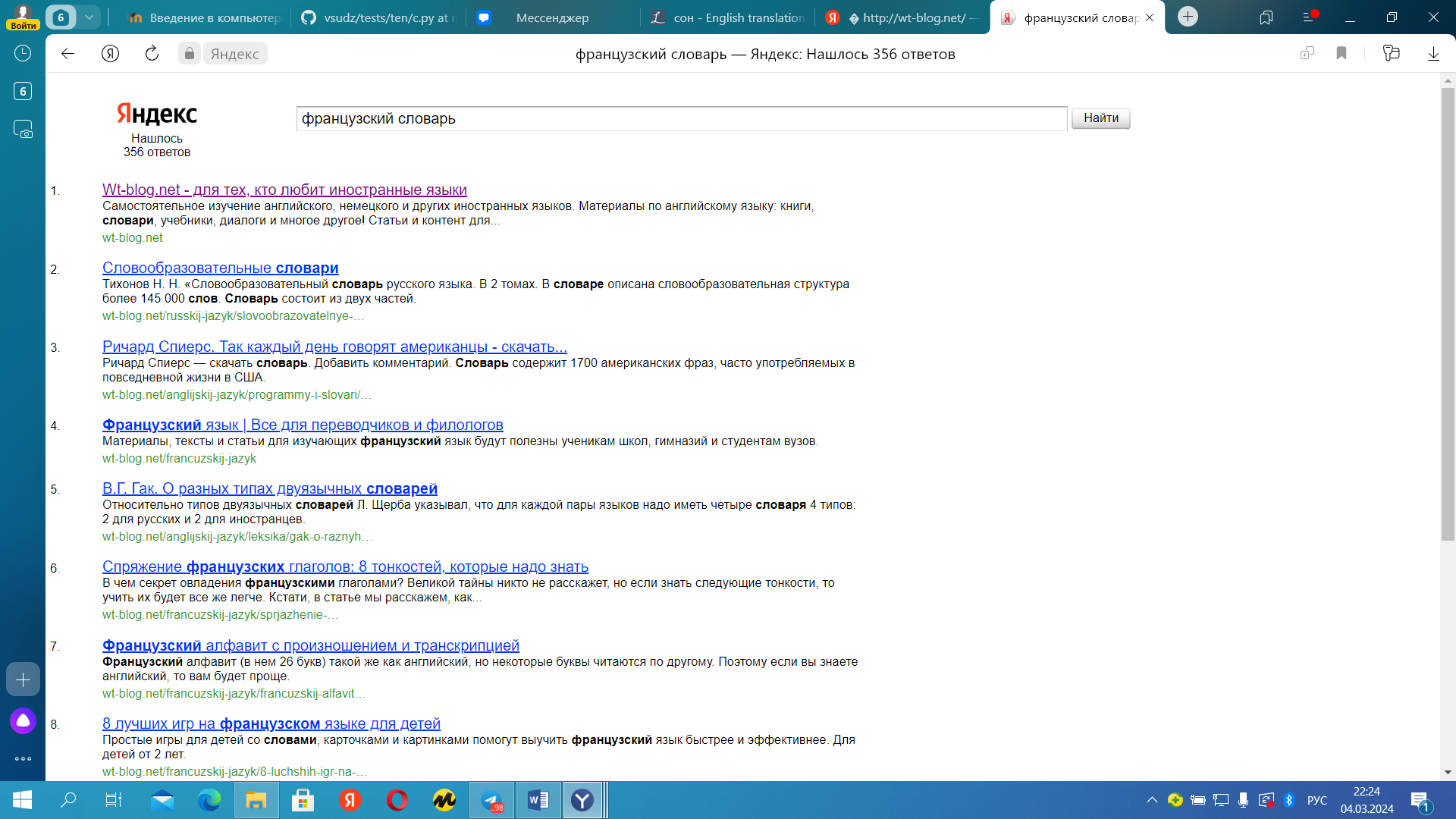
1. Не всегда полная или достоверная информация.

2. Ограниченный выбор тем или периодичность обновления контента.

3. Недостаток интерактивности или возможности задать вопросы автору блога.

Похожие ресурсы на информационные блоги об изучении иностранных языков могут включать другие блоги, форумы, сообщества изучающих языки, онлайн-курсы и ресурсы для самостоятельного обучения языкам.





Действительно хороший сайт для тех, кто хочет изучать языки с помощью словарей. Сайт выдает словари языков, которые мы вводим в запрос.

- Linguee

Linguee - это онлайн-словарь и поисковая система, которая предоставляет переводы и примеры употребления слов и фраз на различных языках. В Linguee также представлена информация о контексте использования слов и выражений, что помогает лучше понять и запомнить их значения.

Для чего можно использовать Linguee:

1. Поиск перевода слов и фраз на различные языки.

2. Поиск контекстуальных примеров использования слов и выражений.

3. Проверка правильности перевода или использования слов в контексте.

4. Изучение новых слов и фраз на иностранных языках.

Для производства запросов в Linguee можно ввести слово или фразу для перевода или поиска контекстуальных примеров. Синтаксис запроса может включать ключевые слова, фразы в кавычках для поиска точного выражения, а также операторы логических связей (AND, OR) для уточнения запроса.

Преимущества Linguee:

1. Большой выбор языков и качественные переводы.

2. Контекстуальные примеры использования слов.

3. Удобный и интуитивно понятный интерфейс.

4. Возможность сравнивать переводы от разных источников.

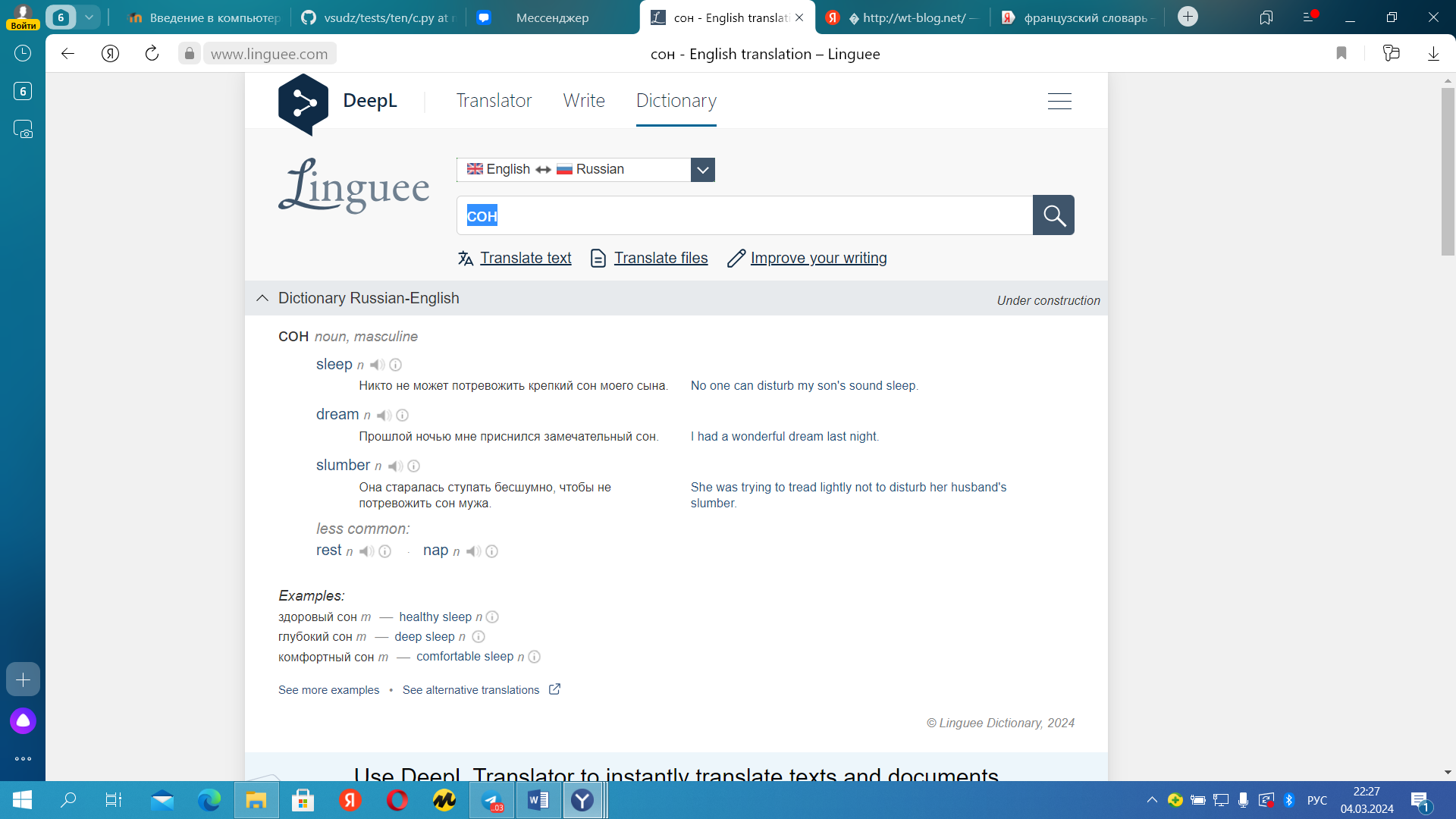
Недостатки Linguee:

1. Не всегда точные или полные переводы.

2. Ограниченный набор языков или тематик.

3. Не всегда доступны все возможные переводы слова или фразы.

Похожие ресурсы на Linguee могут включать другие онлайн-словари, переводчики, ресурсы для изучения языков, такие как Google Translate, Reverso Context, WordReference и др.



Хороший сайт переводчик слов на разные языки.

**Задание 2.**

**Вопросы для обсуждения:**

**1)** Корпус текстов - это вид корпуса данных, единицами которого являются тексты или их достаточно значительные фрагменты, включающие, например, какие-то полные фрагменты макроструктуры текстов данной проблемной области.

**2)** Отбор текстов для корпуса зависит от целей и задач исследования. Принципы отбора могут включать в себя следующие критерии:

1. Репрезентативность: тексты должны быть репрезентативными для изучаемого языка или языковой области.

2. Разнообразие: тексты должны представлять различные жанры (публицистика, художественная литература, научные тексты и т.д.), стили (официальный, разговорный и т.д.) и тематики (политика, экономика, культура и т.д.).

3. Объемность: корпус должен содержать достаточное количество текстов для проведения статистических анализов и обобщений.

4. Контролируемость: отобранные тексты должны быть легко доступными и контролируемыми для исследователя.

5. Стандартизация: тексты должны быть стандартизированными по формату и кодировке для удобства обработки и анализа.

Брауновский корпус - один из первых и наиболее известных корпусов английского языка, который был создан в середине 20 века. Принципы отбора текстов для Брауновского корпуса включали в себя следующие критерии:

1. Репрезентативность: тексты были отобраны таким образом, чтобы они отражали разнообразие жанров, стилей и тематик английского языка. В корпус вошли тексты из различных источников - от газет и журналов до художественной литературы и научных статей.

2. Разнообразие: Брауновский корпус содержал тексты различных жанров (публицистика, художественная литература, научные тексты и т.д.), стилей (официальный, разговорный и т.д.) и тематик (политика, экономика, культура и т.д.).

3. Объемность: Брауновский корпус включал более миллиона слов, что обеспечивало достаточное количество текстов для проведения статистических анализов и обобщений.

**3)** Корпусы языка могут быть классифицированы по различным критериям, таким как:

1. По объему данных:

- Маленькие корпусы: содержат относительно небольшое количество текстовых данных, обычно используются для специфических исследований.

- Большие корпусы: включают в себя огромное количество текстов, что позволяет проводить более обширные исследования и анализы.

2. По типу текстов:

- Общие корпусы: содержат разнообразные тексты из различных источников и жанров.

- Специализированные корпусы: содержат тексты из определенной области знаний или жанра (например, медицинские тексты, юридические документы и т.д.).

3. По источнику данных:

- Собственные корпусы: созданные исследователями или организациями для конкретных целей.

- Открытые корпусы: доступные для общего пользования и собранные из различных источников, таких как интернет, газеты, книги и т.д.

4. По языковой паре:

- Монолингвальные корпусы: содержат тексты только на одном языке.

- Параллельные корпусы: содержат тексты на двух или более языках, соответствующие друг другу.

5. По структуре данных:

- Текстовые корпусы: содержат только текстовую информацию.

- Мультимодальные корпусы: включают в себя различные типы данных, такие как текст, звук, видео и т.д.

6. По цели использования:

- Обучающие корпусы: используются для обучения и разработки моделей машинного обучения.

- Исследовательские корпусы: созданы для лингвистических исследований и анализа языка.

1. Исследовательский корпус: Этот тип корпуса создается с целью исследования конкретных языковых явлений или языковых особенностей. Исследовательские корпусы могут быть собраны для проведения лингвистических исследований, статистических анализов или проверки гипотез. Они обычно могут быть специализированными и ориентированы на определенные языковые явления.

2. Статический корпус: Этот тип корпуса представляет собой набор текстов, который не изменяется или обновляется после создания. Статические корпусы могут использоваться для анализа определенного периода времени или для сравнения изменений в языке на протяжении времени.

3. Параллельный корпус: Параллельные корпуса содержат тексты на двух или более языках, которые соответствуют друг другу по содержанию. Такие корпуса используются для машинного перевода, сравнительного анализа языков, изучения перевода и других задач.

**4)** Для выполнения данного задания я решила выбрать Национальный корпус русского языка.

Национальный корпус русского языка проектировался в начале 2000 годов как корпус современного русского языка по образцу больших национальных корпусов объемом **100 млн** словоупотреблений.

Виды корпусов Национального корпуса русского языка:

1. **Глубоко аннотированный (синтаксический) корпус**. Содержит тексты с морфо-синтаксической разметкой.
2. **Корпус параллельных текстов**. В нём тексту на русском языке сопоставлен перевод на другой язык или наоборот.
3. **Основной корпус текстов**. Представляет русский литературный язык и включает современные письменные тексты и ранние тексты.
4. **Корпус диалектных текстов**. Включает записи диалектной речи в орфографии, приближенной к стандартной.
5. **Акцентологический корпус**. Содержит тексты, несущие информацию об истории русского ударения.
6. **Газетный корпус**. Охватывает статьи из средств массовой информации 2000-х годов.
7. **Корпус поэтических текстов**. Включает стихотворные произведения от XVIII в. до современности.
8. **Мультимедийный корпус**. Содержит фрагменты кинофильмов 1930–2000-х годов.
9. **Корпус устной речи**. Включает расшифровки магнитофонных записей публичной и частной устной речи, а также транскрипты кинофильмов.
10. **Обучающий корпус русского языка**. Ориентирован на преподавание русского языка в школе.

**5)** Глоссарий по теме «Корпусная лингвистика»:

1. Конкорданс:

- Конкорданс представляет собой список контекстов, в которых встречается определенное слово или выражение в корпусе текстов. Конкордансы используются для анализа употребления слова в различных контекстах.

2. Рандомизация:

- Рандомизация в корпусной лингвистике относится к процессу случайного выбора текстов или слов из корпуса для обеспечения репрезентативности выборки и уменьшения искажений результатов исследования.

3. Коллокация:

- Коллокация - это устойчивое сочетание слов, которые часто встречаются вместе и образуют единое значение. Изучение коллокаций помогает понять связи между словами в языке.

4. Подмассив:

- Подмассив - это часть корпуса текстов, выделенная для конкретного исследования или анализа. Подмассивы позволяют сузить область исследования и сосредоточиться на конкретных языковых явлениях.

5. Парсинг:

- Парсинг - это процесс анализа грамматической структуры предложения с целью определения зависимостей между словами и их ролями в предложении. Парсинг используется для синтаксического анализа текстов.

6. Лемматизация:

- Лемматизация - это процесс приведения словоформы к начальной форме (лемме) с целью унификации словоупотреблений. Лемматизация помогает учитывать различные формы слов как одно и то же слово при анализе текстов.

7. Корпус-менеджер:

- Корпус-менеджер - это программное обеспечение, предназначенное для работы с корпусами текстов, включая их создание, хранение, поиск, анализ и аннотацию. Корпус-менеджеры облегчают доступ к данным и проведение лингвистических исследований.

**6)** Сетевые ресурсы по теме корпусная лингвистика представляют собой онлайн-платформы, базы данных или инструменты, предназначенные для работы с текстовыми корпусами (крупными собраниями текстов) с целью исследования языка. Общие черты таких ресурсов включают:

1. Доступ к большим объемам текстов: Сетевые ресурсы по корпусной лингвистике обычно содержат обширные коллекции текстов на различных языках и различных жанрах, что позволяет исследователям проводить анализ языка на основе большого объема данных.

2. Поиск и фильтрация данных: Пользователи могут осуществлять поиск по корпусу текстов, применять различные фильтры (по автору, дате, жанру и т.д.) для уточнения результатов и нахождения нужной информации.

3. Инструменты для анализа текста: Некоторые сетевые ресурсы предоставляют инструменты для анализа текста, такие как частотный анализ слов, анализ коллокаций, определение частей речи и т.д., что помогает исследователям изучать языковые явления.

4. Возможность сравнения и сопоставления данных: Некоторые ресурсы позволяют сравнивать различные тексты или языки, что полезно для сопоставительного анализа и изучения языковых различий.

5. Общедоступность: Большинство сетевых ресурсов по корпусной лингвистике доступны для использования бесплатно или при условии регистрации, что делает их доступными для широкого круга пользователей, включая студентов, преподавателей и исследователей.

Некоторые популярные сетевые ресурсы по корпусной лингвистике включают Sketch Engine, Corpus of Contemporary American English (COCA), British National Corpus (BNC), Google Ngram Viewer и другие.

**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.

**Уровень 2.**

**Задание 2.**

|  |  |
| --- | --- |
| В чём сущность эмпирической методологии в лингвистическом анализе? | Сущность этой методологии заключается в том, что она стремится к объективному и научному изучению языка на основе наблюдений, экспериментов и анализа эмпирических данных. |
| Дайте определение понятию коллокация | Сущность этой методологии заключается в том, что она стремится к объективному и научному изучению языка на основе наблюдений, экспериментов и анализа эмпирических данных. |
| Сформулируйте закон Ципфа. | Это  [эмпирическая закономерность](https://ru.wikipedia.org/wiki/%D0%AD%D0%BC%D0%BF%D0%B8%D1%80%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D0%B7%D0%B0%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%80%D0%BD%D0%BE%D1%81%D1%82%D1%8C) распределения [частотности](https://ru.wikipedia.org/wiki/%D0%A7%D0%B0%D1%81%D1%82%D0%BE%D1%82%D0%BD%D0%BE%D1%81%D1%82%D1%8C) [слов](https://ru.wikipedia.org/wiki/%D0%A1%D0%BB%D0%BE%D0%B2%D0%BE) [естественного языка](https://ru.wikipedia.org/wiki/%D0%95%D1%81%D1%82%D0%B5%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA): если все слова [языка](https://ru.wikipedia.org/wiki/%D0%AF%D0%B7%D1%8B%D0%BA) (или просто достаточно длинного [текста](https://ru.wikipedia.org/wiki/%D0%A2%D0%B5%D0%BA%D1%81%D1%82)) упорядочить по убыванию [частотности](https://ru.wikipedia.org/wiki/%D0%A7%D0%B0%D1%81%D1%82%D0%BE%D1%82%D0%BD%D0%BE%D1%81%D1%82%D1%8C) их использования, то частотность *n*-го слова в таком списке окажется приблизительно [обратно пропорциональной](https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%BE%D0%BF%D0%BE%D1%80%D1%86%D0%B8%D0%BE%D0%BD%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE%D1%81%D1%82%D1%8C) его порядковому номеру *n* (так называемому *рангу* этого слова, см. [шкала порядка](https://ru.wikipedia.org/wiki/%D0%A8%D0%BA%D0%B0%D0%BB%D0%B0)). Например, второе по используемости слово встречается примерно в два раза реже, чем первое, третье — в три раза реже, чем первое, и так далее. |
| Дайте определениепонятию дискурсивной просодии | Использование интонации и других звуковых характеристик речи для организации дискурса или разговора.Д |
| Дайте определение понятию семантической просодии | Использование интонации, тонов и ритма в речи для передачи смысла и эмоциональной окраски высказывания. |
| В чём сущность гипотизы общего ядра? | Существует некоторое общее ядро лексических единиц или концептов, которые являются универсальными для всех языков. |
| Что такое дерево зависимостей? | Наиболее наглядный и наиболее распространенный способ представления синтаксической структуры предложения. |
| Какие основные сферы применения корпусов? | Разработка словарей, анализ текстов. |
| Какие слова являются синонимами термина аннотация? | Сочинение, изложение. |
| Дайте определение понятию репрезентативность корпуса? | Свойство корпуса, заключающееся в статистически достоверном представлении языка или его части и достигаемое за счет необходимого объема и жанрового разнообразия текстов. |
| Дайте определение понятию кластерный анализ? | Это многомерная статистическая процедура, которая выполняет сбор данных, содержащих информацию о выборке объектов, и затем упорядочивает объекты в сравнительно однородные группы. |
| Что такое конкорданс? | Это упорядоченный список всех употреблений заданных слов (или выражений) в контексте со ссылками на источник в заданном тексте или в работах отдельных авторов. |
| Что такое коркондансер? | Это специальная программы для автоматического составления конкордансов. |
| Дайте определение понятию дистрибуции? | Распределение языковых единиц в потоке речи или в системе языка, их закономерные сочетания друг с другом. |
| Дайте определение понятию частотность? | Отношение количества экземпляров данного объекта полному количеству экземпляров всех объектов в данном множестве. |
| Дайте определение понятию лемма, лемматизация? | Лемма - это основная (нормализованная) форма слова, которая представляет собой его словарную форму.  Лемматизация - это процесс приведения слова к его лемме. |
| Дайте определение понятию лексическая плотность? | Измеряет структуру и сложность человеческого общения на том или ином языке. |
| Что такое морфологический анализатор? | Это инструмент, определяющий грамматические характеристики слова. |
| Дайте определение понятию сегментация корпуса? | Процесс разделения текстового корпуса на отдельные единицы, такие как предложения, абзацы или слова. |

**Тест.**

1. В чём сущность эмпирической методологии в лингвистическом анализе?

 Сущность этой методологии заключается в том, что она стремится только к объективному изучению языка на основе наблюдений, экспериментов и анализа эмпирических данных.

 Сущность этой методологии заключается в том, что она стремится к объективному и научному изучению языка на основе гипотиз.

Сущность этой методологии заключается в том, что она стремится к объективному и научному изучению языка на основе наблюдений, экспериментов и анализа эмпирических данных.

2. Какое верное определение понятия коллокация?

 Это совсем не устойчивые сочетания слов, которые часто встречаются вместе и образуют единую лексическую или грамматическую единицу.

 Это устойчивые сочетания слов, которые часто встречаются вместе и образуют единую лексическую или грамматическую единицу.

 Это устойчивые сочетания слов, которые редко встречаются вместе и образуют единую лексическую или грамматическую единицу.

3. Сформулируйте закон Ципфа.

Это  [эмпирическая закономерность](https://ru.wikipedia.org/wiki/%D0%AD%D0%BC%D0%BF%D0%B8%D1%80%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D0%B7%D0%B0%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%80%D0%BD%D0%BE%D1%81%D1%82%D1%8C) распределения [частотности](https://ru.wikipedia.org/wiki/%D0%A7%D0%B0%D1%81%D1%82%D0%BE%D1%82%D0%BD%D0%BE%D1%81%D1%82%D1%8C) [слов](https://ru.wikipedia.org/wiki/%D0%A1%D0%BB%D0%BE%D0%B2%D0%BE) [естественного языка](https://ru.wikipedia.org/wiki/%D0%95%D1%81%D1%82%D0%B5%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA): если все слова [языка](https://ru.wikipedia.org/wiki/%D0%AF%D0%B7%D1%8B%D0%BA) (или просто достаточно длинного [текста](https://ru.wikipedia.org/wiki/%D0%A2%D0%B5%D0%BA%D1%81%D1%82)) упорядочить по убыванию [частотности](https://ru.wikipedia.org/wiki/%D0%A7%D0%B0%D1%81%D1%82%D0%BE%D1%82%D0%BD%D0%BE%D1%81%D1%82%D1%8C) их использования, то частотность *n*-го слова в таком списке окажется приблизительно [обратно пропорциональной](https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%BE%D0%BF%D0%BE%D1%80%D1%86%D0%B8%D0%BE%D0%BD%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE%D1%81%D1%82%D1%8C) его порядковому номеру *n* (так называемому *рангу* этого слова, см. [шкала порядка](https://ru.wikipedia.org/wiki/%D0%A8%D0%BA%D0%B0%D0%BB%D0%B0)). Например, второе по используемости слово встречается примерно в два раза реже, чем первое, третье — в три раза реже, чем первое, и так далее.

 Математическая формула для вычисления частоты слов

 Правило определения синонимов

4. Какое верное определение понятия дискурсивной просодии?

 Анализ семантической структуры предложений

 Использование интонации, тонов и ритма в речи для передачи смысла и эмоциональной окраски высказывания.

Использование интонации и других звуковых характеристик речи для организации дискурса или разговора.

5. Какое верное определение понятия семантической просодии?

 Анализ ритма и интонации в речи

Использование интонации, тонов и ритма в речи для передачи смысла и эмоциональной окраски высказывания.

 Изучение структуры предложений

6. В чём сущность гипотизы общего ядра?

 Существует некоторое общее ядро лексических единиц или концептов, которые являются универсальными для всех языков.

 Идея о наличии общего языкового корня у разных языков

 Гипотеза о существовании универсальных языковых категорий

7. Что такое дерево зависимостей?

 Метод анализа морфологии слов

 Способ классификации слов по частоте употребления

 Наиболее наглядный и наиболее распространенный способ представления синтаксической структуры предложения.

8. Какие основные сферы применения корпусов?

 Медицина, искусственный интеллект, физика

Разработка словарей, анализ текстов.

 История, археология, антропология

9. Какие слова являются синонимами термина аннотация?

Сочинение, изложение.

Диктовка текста.

Копирование текста.

10. Какое верное определение понятия репрезентативность корпуса?

 Объем текстов, включенных в корпус для анализа

 Методика оценки достоверности данных в корпусе

 Свойство корпуса, заключающееся в статистически достоверном представлении языка или его части и достигаемое за счет необходимого объема и жанрового разнообразия текстов.

11. Какое верное определение понятия кластерный анализ?

 Исследование структуры предложений в тексте

Это многомерная статистическая процедура, которая выполняет сбор данных, содержащих информацию о выборке объектов, и затем упорядочивает объекты в сравнительно однородные группы.

 Оценка частоты употребления лексических единиц

12. Что такое конкорданс?

 Это упорядоченный список всех употреблений заданных слов (или выражений) в контексте со ссылками на источник в заданном тексте или в работах отдельных авторов.

 Метод исследования морфологии слов

 Теория о структуре предложений

13. Что такое коркондансер?

 Устройство для анализа семантики предложений

 Специалист по лингвистике корпусов

 Это специальная программы для автоматического составления конкордансов.

14. Какое верное определение понятия дистрибуции?

 Анализ морфологических характеристик слов

 Распределение языковых единиц в потоке речи или в системе языка, их закономерные сочетания друг с другом.

 Структура предложений в тексте

15. Какое верное определение понятия частотность?

 Отношение количества экземпляров данного объекта полному количеству экземпляров всех объектов в данном множестве.

 Семантическая связь между словами

 Грамматическая категория слова

16. Какое верное определение понятия лемма, лемматизация?

 Анализ морфологических особенностей слова

 Оценка частоты употребления слов в тексте

 Лемма - это основная (нормализованная) форма слова, которая представляет собой его словарную форму.

Лемматизация - это процесс приведения слова к его лемме.

17. Какое верное определение понятия лексическая плотность?

 Исследование структуры предложений на уровне морфологии

 Измеряет структуру и сложность человеческого общения на том или ином языке

 Оценка степени формальной сложности текста

18. Что такое морфологический анализатор?

 Это инструмент, определяющий грамматические характеристики слова.

 Методика анализа синтаксической структуры предложения

 Специалист по изучению морфологии языка

19. Какое верное определение понятия сегментация корпуса?

 Объединение различных корпусов для создания большего объема данных

 Преобразование текстовых данных в числовые значения

 Процесс разделения текстового корпуса на отдельные единицы, такие как предложения, абзацы или слова

**Задание 3.**

**Тест 1.**

1. Тексты, составляющие лингвистический корпус,

организованы для удобства чтения.

сконструированы лингвистами.

не включают речь необразованных носителей языка.

размечены лингвистами и программистами.

2. В лингвистический корпус могут входить

тексты электронной коммуникации.

записи устной речи детей.

транскрипции теле- и радиопередач.

произведения художественной литературы.

3. Для того, чтобы электронная коллекция текстов могла считаться корпусом,

в ней не должно быть текстов, посвященных неактуальным для общества проблемам.

она должна отражать состояние языка в определенный период времени.

она должны быть размечена.

в ней должны быть исправлены ошибки носителей языка.

4. Обязательными свойствами хорошего лингвистического корпуса являются

представительность.

присутствие разметки.

снятая грамматическая омонимия.

баланс между разными типами текстов.

5. Создавая репрезентативный корпус, лингвисты должны ответить на вопросы:

Какие тексты считать значимыми для данной культуры?

Отражают ли оригинальные тексты художественной литературы состояние языка эпохи?

Можно ли, изучая язык периода или жанра, опираться на переводные тексты?

Какое количество текстов будет достаточным для демонстрации состояния языка конкретного периода?

6. Интерпретация корпусных данных невозможна, если исследователю неизвестен \_\_\_интерфейс\_\_\_\_\_ корпуса, по которому велся поиск.

7. Согласно материалам лекции, основная часть Национального корпуса русского языка включает \_\_\_\_миллиарды\_\_\_\_ слов.

8. Количество найденных результатов обозначается в НКРЯ как количество «\_\_\_\_найденных\_\_\_\_».

9. Нормализация значений обычно предполагает вычисление отношения количества вхождений к объему корпуса, умноженного на \_\_\_\_общий объём\_\_\_\_.

10. Пропорциональное представление в корпусе текстов разных типов называется «\_\_\_\_балансирование\_\_\_\_» корпуса.

11. Что означает понятие «накопительная сбалансированность»?

В большом корпусе в некоторых случаях можно пренебречь представлением текстов разных типов в соотношениях, сопоставимых с соотношениями в естественном языке.

С увеличением числа текстов, входящих в корпус, увеличивается необходимость разнообразить состав включаемых в него документов.

Лингвистический корпус может быть относительно небольшим, важно, чтобы в нем были представлены разные подъязыки в пропорциях, соответствующих аналогичным пропорциям в естественном языке.

Понятия «сбалансированность» и «репрезентативность» никак не связаны; хороший корпус должен обладать обеими характеристиками.

12. Что отличает разметку Обучающего корпуса от аннотации других частей НКРЯ?

В ней присутствует синтаксическая разметка.

Она проведена в соответствии со школьной грамматикой.

Она позволяет вести поиск по речевым действиям.

Она целиком произведена вручную.

13. Чем отличается лингвистический корпус от электронного словаря?

Электронный словарь не предлагает контекстов употребления слова.

В лингвистическом корпусе языковая норма отражена только опосредованно.

Материалы лингвистического корпуса не дают представления о том, насколько слово распространено в языке.

В электронном словаре могут содержаться сведения о семантике слова.

14. Какой вид поиска невозможен для пользователя в поисковых системах Яндекс и Google?

Поиск в пределах одного сайта.

Поиск точных форм.

Поиск словоформ одной лексемы.

Поиск смысловых категорий.

15. На каком ресурсе можно получить информацию о частотности употребления слова в конкретный период времени?

Яндекс Переводчик

Google Books Ngram Viewer

ABBYY LINGVO

Google Translator Toolkit

**Тест 2.**

1. Утверждение «Корпус – это интерпретированный текст» означает, что

при помощи разметки создатели корпуса передают пользователю свои теоретические установки.

прежде, чем включать тексты в корпус, лингвисты тщательно изучают их и делают выводы об их содержании.

начиная работу в корпусе, пользователь по умолчанию соглашается принять на веру принципы, положенные в основу аннотации.

каждое языковое явление должно быть интерпретировано создателями корпуса как правильное или неправильное и соответствующим образом помечено.

2. Аннотирование – это

обработка текста в целях облегчения его изучения.

необходимый этап создания лингвистического корпуса.

присвоение каждому элементу текста особого тэга.

сведения, размещенные на главной странице корпуса.

3. Утверждение «Разметка не несет новой информации» означает, что

разметчики описывают характеристики, присущие элементам текста.

принципы разметки – это внутренняя информация, важная только для создателей корпуса.

теоретические установки, положенные в основу аннотации, должны быть максимально нейтральными.

разметка не выводится на экраны пользователей корпуса.

4. Автоматическая разметка

начинается с деления текста на минимальные фрагменты.

невозможна на уровне семантики.

предполагает отнесение словоформы к начальной форме.

позволяет быстро провести аннотирование по частям речи.

5. Экстралингвистическая разметка

предполагает демонстрацию сведений об авторе.

не затрагивает содержания текста.

позволяет вести поиск в отобранной по социолингвистическим параметрам группе текстов.

в НКРЯ выражается в виде «паспорта текста».

6. Морфологическая разметка в НКРЯ

произведена с помощью компьютерных программ.

началась с выработки командой лингвистов стандартных решений в спорных ситуациях.

единообразна во всех частях НКРЯ.

предполагает присвоение тэгов каждой единице текста.

7. Синтаксическая разметка в НКРЯ НЕ

проведена в соответствии с традиционным выделением членов предложения.

графически представлена в виде дерева зависимостей.

проведена компьютерной программой, затем доработана вручную.

охватывает все части НКРЯ.

8. Семантическая разметка в НКРЯ

помечает семантические признаки грамматических разрядов.

помечает семантические признаки словообразовательных показателей.

не затрагивает собственно семантические признаки.

предполагает присвоение одной лемме одного семантического тэга.

9. Утверждение «Мультимодальный корпус требует нескольких уровней разметки» предполагает, что

для корпуса МУРКО в составе НКРЯ были разработаны дополнительные принципы описания текстов, например, создана разметка значений жестов.

элементы корпуса, представляющего визуальные образы, должны быть описаны иначе, чем элементы письменного или аудио текста.

мультимодальные корпуса должны быть частично закрытыми из соображений этики.

разметка по цели высказывания или по дискурсивным элементам не нужна в обычных текстовых корпусах.

10. В корпусе с неснятой грамматической омонимией словоформа «уже» будет отнесена к

одной лемме.

двум леммам.

трем леммам.

четырем леммам.

11. В качестве минимального элемента разметки в НКРЯ принята \_\_\_\_морфема\_\_\_\_, то есть графическая единица, не обладающая, в отличии от лексемы, семантическими характеристиками.

12. Производимый обычно вручную процесс снятия омонимии носит название «\_\_\_\_лемматизация\_\_\_\_».

13. \_\_\_Матрица\_\_\_\_\_ – это компьютерная программа, производящая синтаксическую разметку.

14. Многие автоматические анализаторы морфологии русского языка основаны на данных «\_\_\_морфологического\_\_\_\_\_ словаря» А. Зализняка.

15. Источником автоматической семантической разметки часто служат компьютерные \_\_\_\_корпуса\_\_\_\_.

**Уровень 3**

**Задание 2.**

**1.** Для исследования я решила выбрать тему «Лингвистические особенности написания художественных произведений».

Я взяла отрывок из текста «Алиса в стране чудес.

Алисе начинало надоедать сидеть рядом с сестрой на берегу и ничего не делать: раз или два она заглядывала в книгу, которую читала сестра, но в ней не было ни картинок, ни разговоров, «а что толку от книги, — думала Алиса, — без картинок и разговоров?». И вот она раздумывала (как могла, потому что жаркий день сделал ее очень сонной и глупой), стоит ли удовольствие сделать цепочку из маргариток, чтобы вставать и собирать их, как вдруг рядом с ней пробежал Белый Кролик с розовыми глазами. В этом не было ничего примечательного; Алиса также не подумала, что это так уж необычно, когда услышала, как Кролик сказал себе: «О боже! О боже! Я опоздаю!» (потом, когда она обдумала это, ей пришло в голову, что она должна была удивиться этому, но в то время все это казалось вполне естественным); но когда Кролик действительно достал часы из жилетного кармана, посмотрел на них, а затем поспешил дальше, Алиса вскочила на ноги, потому что ей пришло в голову, что она никогда раньше не видела кролика ни с жилетным карманом, ни с часами, которые можно было бы из него достать, и, сгорая от любопытства, она побежала по полю вслед за ним, и, к счастью, как раз вовремя, чтобы увидеть, как он спускается в большую кроличью нору под живой изгородью. Еще через мгновение Алиса спустилась за ним, ни разу не задумавшись о том, как ей выбраться обратно.

В данном тексте преобладают глаголы и местоимения. Глаголы помогают передать динамику событий и создать образы воображаемых сцен. Местоимения помогают определить субъекты и объекты действий. В художественных текстах основное внимание уделяется действиям и внутреннему мироощущению персонажей. Также в художественных текстах часто присутсвуют диалоги, они помогают передать характеры персонажей, их отношения друг к другу, а также развивать сюжет. Художественные произведения обычно имеют определенную структуру и композицию, которые помогают автору развивать сюжет, создавать напряжение и поддерживать интерес читателя.

**2.** Частота употребления имен существительных в притяжательном падеже (baby’s bottle, car’s wheel, government’s decision) в английской версии языка составляет 172, в британской версии языка 164. Почти одна и та же частота, потому что эти выражения являются распространёнными в обоих языках.

**3.** Из анализа корпусов текстов видно, что обычно рефлексивный глагол с местоимением self употребляется в следующих форматах:

- Глагол + self: "He prides himself on his achievements."

- Местоимение + глагол + self: "My sister always blames herself for everything."

- Местоимение + глагол + дополнение + self: "She found herself in a difficult situation."

Таким образом, наиболее часто встречающиеся варианты сочетаемости рефлексивного глагола с местоимением self включают в себя указанную выше структуру словосочетания.

**4.** Частотность сочетаемости глагола help с чистым инфинитивом для британского варианта английского языка составляет примерно 28,41. Это значит, что данная конструкция используется весьма часто.

**5**. Например, возьмем пару синонимических прилагательных beautiful и gorgeous и проанализируем корпуса текстов.

Прилагательное **gorgeous** имеет такие значения: очень красивый, привлекательный, приятный, дарящий наслаждение.

Прилагательное **beautiful** имеет смысловой оттенок «доставляющий эстетическое наслаждение». Это может быть женское лицо, фигура или вещь, пейзаж, музыка и т.д.

Можем заметить, что прилагательные имеют одно значение, но употребляются в разных ситуациях. Частота употребления слова beautiful больше, чем слова **gorgeous.**

**6. - Глагол "eat": в конструкциях с глаголом "eat" наиболее часто употребляемыми объектами являются еда (food), фрукты (fruits), овощи (vegetables), мясо (meat), хлеб (bread) и другие продукты питания.**

**- Глагол "drink": в случае с глаголом "drink" часто употребляемыми объектами являются напитки (drinks) как алкогольные, так и безалкогольные, такие как вода (water), чай (tea), кофе (coffee), соки (juices) и т.д.**

**7.** **Для разграничения синтаксических синонимов на основе лингвистических корпусов можно использовать следующие подходы:**

**1. Контекстуальный анализ:**

**- Изучение контекста употребления слова или выражения в предложениях и текстах из корпусов. Сравнение контекстов, в которых употребляются синонимы, может помочь выявить различия в их использовании.**

**2. Частотный анализ:**

**- Анализ частотности употребления синонимов в различных контекстах может помочь определить предпочтительные сочетания или конструкции для каждого из них.**

**3. Коллокационный анализ:**

**- Изучение типичных коллокаций (словосочетаний) для каждого из синонимов. Различия в типичных сочетаниях могут указывать на специфические контексты использования.**

**4. Семантический анализ:**

**- Исследование семантических оттенков и значений синонимов в контексте их употребления. Определение, какие аспекты значения подчеркиваются при использовании каждого синонима.**

**5. Синтаксический анализ:**

**- Анализ структуры предложений, в которых употребляются синонимы, с целью выявления различий в синтаксической функции или синтаксическом окружении этих слов.**

**8. Использование Present Perfect и Past Simple:**

Present Perfect используется для описания действий или событий, которые произошли в неопределенный момент в прошлом, но имеют отношение к настоящему времени, а Past Simple используется для описания конкретных действий или событий, которые произошли в определенный момент в прошлом.

Примеры синтаксических конструкций:

- Present Perfect (I have lost my keys. (Я потерял ключи.))

- Past Simple (We watched a movie yesterday. (Мы посмотрели фильм вчера.))

В научных текстах чаще используется Past Simple для описания прошлых исследований, результатов и фактов. Present Perfect может использоваться для описания последствий или актуальности полученных результатов.

**9.** - Визуализация текста - изображение либо элементов текста, либо структур, извлечённых из текста, для образовательных или аналитических нужд.

- Визуализация текста может помочь наглядно продемонстрировать закон Ципфа. Например, построение графика с частотой встречаемости слов в тексте по их рангу может показать, насколько точно распределение частотности соответствует закону Ципфа. Если точки на графике образуют прямую линию, это может свидетельствовать о соблюдении закона Ципфа в данном тексте.

**11.** Many Eyes - это онлайн-сервис от IBM, который позволяет пользователям создавать интерактивные визуализации данных, включая текстовую информацию. Система предлагает различные типы визуализации, такие как облако слов, диаграммы, графики и т.д. Пользователи могут загружать свои данные или использовать общедоступные наборы данных для создания визуализаций.

При сравнении Many Eyes с другими ресурсами визуализации текста, можно отметить следующие особенности:

1. Many Eyes предлагает широкий выбор типов визуализации и инструментов для работы с данными, что делает его удобным инструментом для создания разнообразных визуализаций.

2. Некоторые другие ресурсы визуализации текста могут иметь более продвинутые функции и возможности настройки визуализаций. Например, Tableau или Power BI позволяют создавать сложные дашборды и аналитические отчеты на основе текстовой информации.

3. Профессиональные лингвисты могут использовать различные инструменты визуализации текста для анализа языковых данных, выявления закономерностей и трендов. Например, они могут использовать визуализации для исследования структуры текста, анализа частотности слов, выявления ключевых тем и т.д.

4. При выборе системы визуализации текста профессиональные лингвисты могут учитывать не только возможности инструмента, но и его удобство использования, скорость работы, доступность функций анализа данных и т.д.

В целом, сравнение Many Eyes с другими ресурсами визуализации текста может зависеть от конкретных потребностей пользователя и задач анализа текста. Каждый инструмент имеет свои особенности и преимущества, которые можно учитывать при выборе подходящего ресурса для работы с текстовыми данными.

**12.** 1. Лингвистика - наука, которая изучает языковую способность человека в целом, рассматривая каждый конкретный язык, как реализацию уникальной человеческой способности.

2. Компьютерная лингвистика 1 - это лингвистика и есть. Компьютерная лингвистика 2(Так определение назвал сам лектор) - технология методология решения различных полезных задач, которые естественном образом возникли от того, что мы все являемся пользователями компьютеров.

3. Машинный перевод (определение в этом видео не дано, я из собственного опыта дала определение) - перевод, выполненный с помощью компьютера (Например Google переводчик) (это определение я узнала из предыдущей лабораторной).

4. Решение задач на основе обработке ЕЯ должно основываться на таком последовательном анализе языка как системы.

5. (Этапы):

- Модель текста с точки зрения лексический анализа (говорится про письменную речь) - это вещь, которая нигде не описана, но это могут быть разные смайлики, телефонные номера, электронные адреса...

- Морфологический анализ определяется от какой формы слова она определена и приписываются характеристики (например для имя существительного или прилагательного...).

- Синтаксический анализ - это есть структура связи в предложении, которая называется синтаксис.

- Семантический анализ - переход от синтаксической структуры к чему-то более глубокому.

- Прагматический анализ - интерпретация уже то, что получено в результате языковой обработки в контексте ситуации.

6. Треугольник перевода - схема, описывающая автоматическую обработку текста.

7. Неоднозначность в компьютерной лингвистике - пронизывание языка на всех его уровнях.

8. Несеметричность - различные способы кодирования некоторого смысла в разных языках.

9. Избыточность - способы выразить смысл, где в различных языках свой смысл.

10.Конвецианальность - выражение, которое не принято использовать. (Рой собак)

11. Непрозрачность - активное использование сложных средств референтности.

12. Полнота грамматических описаний - степень деятельности и подробности, с которой описаны грамматические описания в данном языке.

13. Региональная норма - то, что является общепринятыми способами номинации в каком-то регионе, за пределами которого она никому не известна.

14. Языковое сообщение - основная единица коммуникации, передаваемая с помощью языка.

15. Статистический подход - метод анализа данных и явлений, основанный на использовании статистических методов для изучения закономерностей, взаимосвязей и тенденций в данных.

16. Информационный анализ - процесс сбора, оценки, интерпретации и представления информации для принятия решения.

17. Лингвистическая структура - организация языковых единиц (слов, фраз, предложений) в рамках конкретного языка.

18. Чистый подход - методы и подходы, которые сосредоточены исключительно на использовании компьютерных алгоритмов и моделей для решения лингвистических задач.

19. Лингвистические подходы - методы анализа и обработки ЕЯ, на основании лингвистических принципах и теориях.

20. Магистральный путь компьютерной лингвистики – соединение того, что получено в ходе статистического анализа и того, что происходит в лингвистики.

21. Машинное обучение - подраздел искусственного интеллекта, который

изучает методы и алгоритмы, позволяющие компьютерным системам

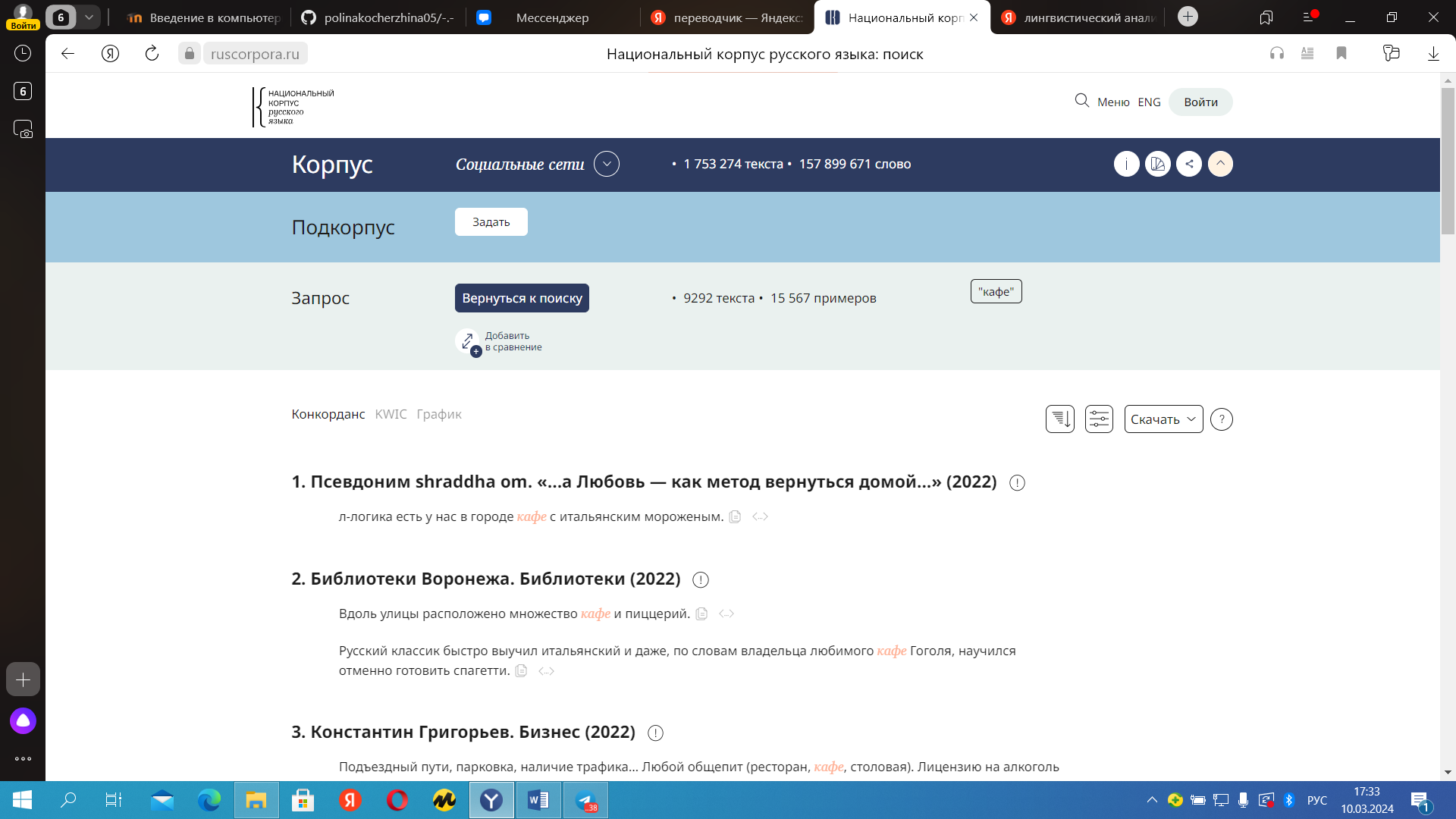
обучаться на основе данных, а не явно программироваться.

22. Контролируемые языки – способ пользователю системы машинного перевода и системе найти друг друга.

После 1:50 пропадает звук в лекции.

**13.** Для выполнения этого задания я решила рассмотреть «Национальный корпус русского языка».

Чтобы провести лингвистический анализ я выбрала корпус социальные сети. Для поиска точных форм я решила взять слова кафе, кино, стул, шкаф. Взяв эти слова я хочу посмотреть на какую тематику будет больше текстов.

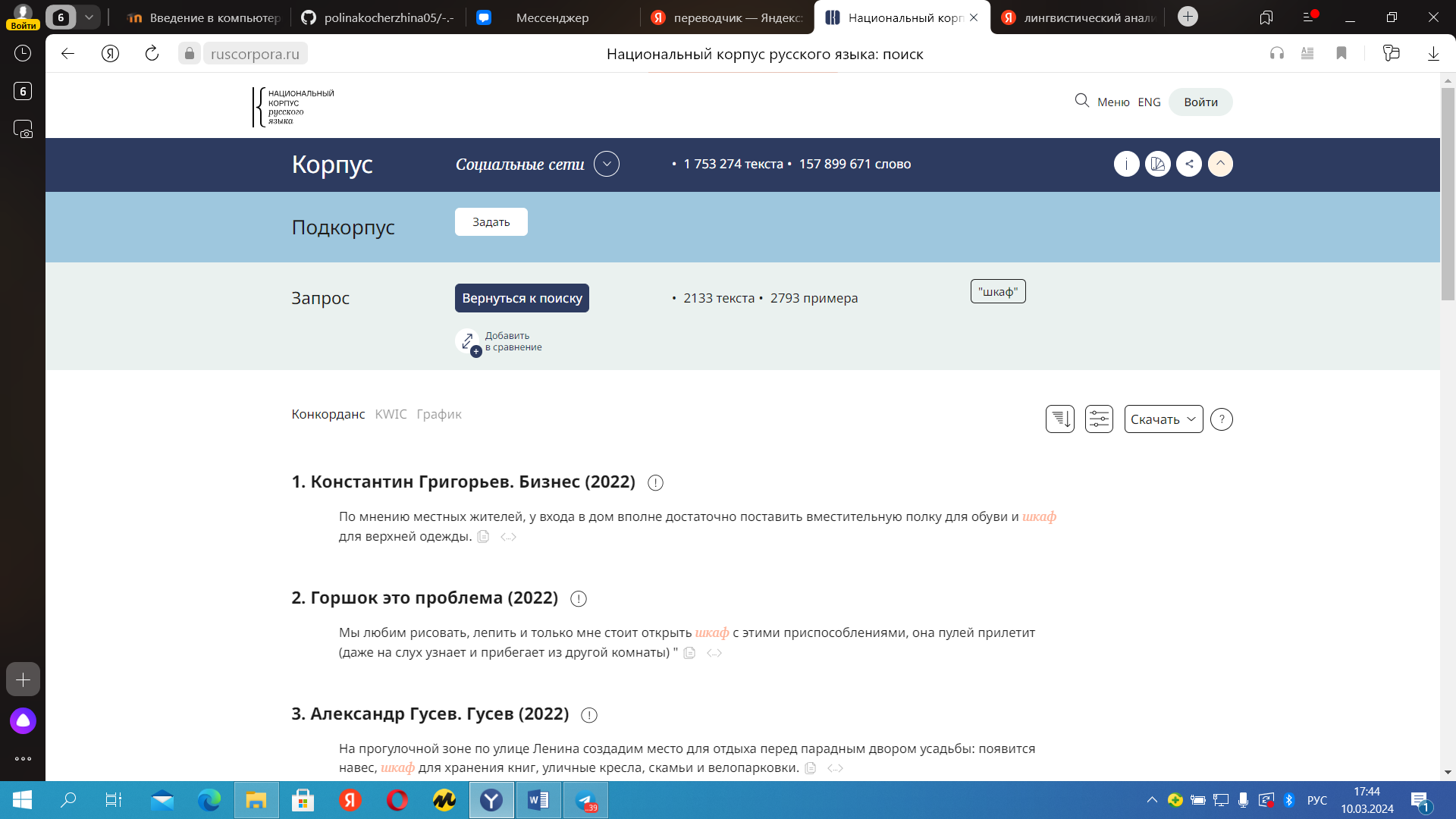


Со словом кафе мне выдало 9 292 текста.

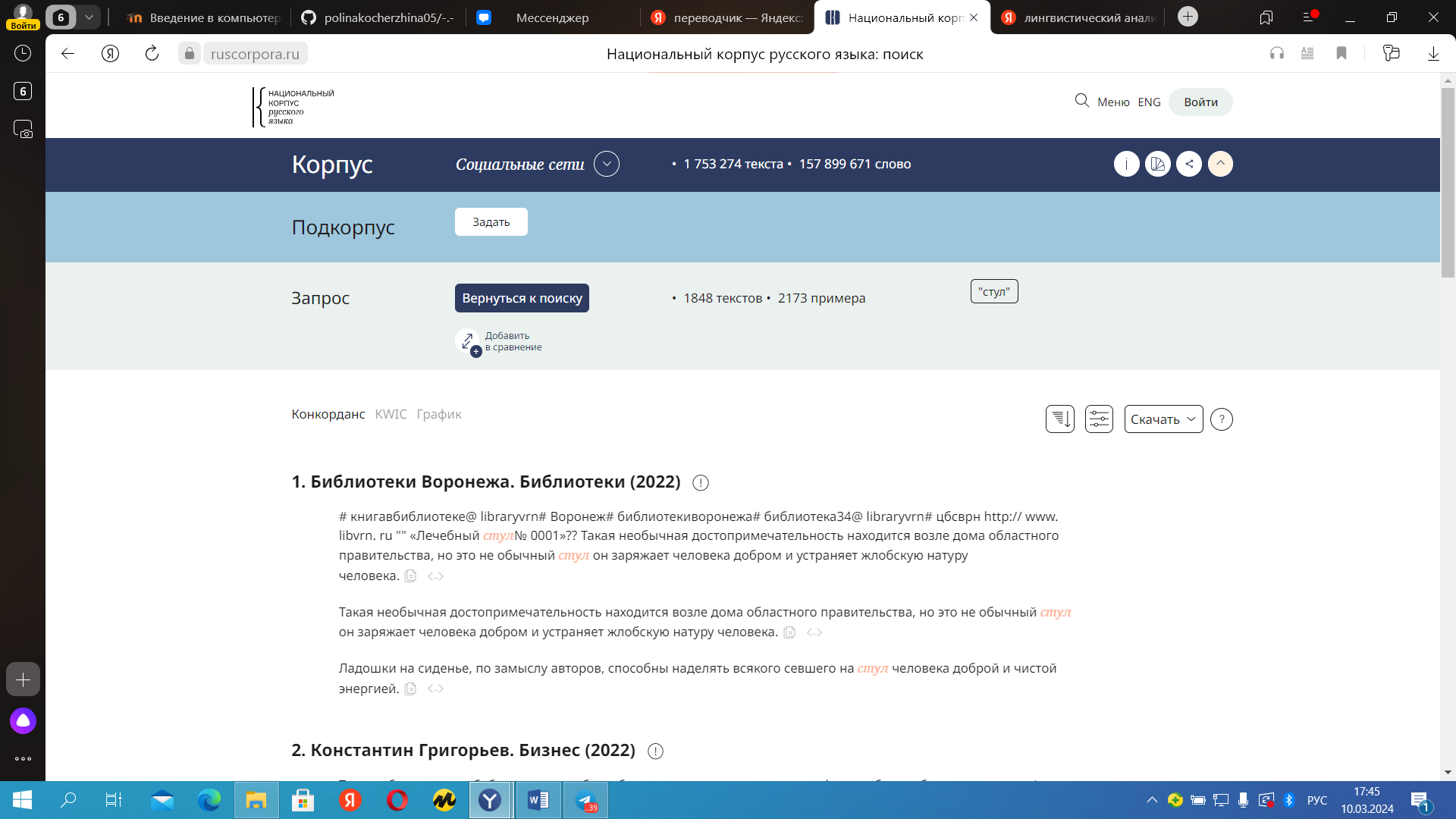
Далее вобьём в поиск слово кино.



Получаем 14 675 текстов.



Со словом шкаф 2 133 текста.



На основе данных запросов можно сделать вывод, что при поиске мест для отдыха текстов выводится в разы больше, чем при поиске мебели. Возможно это происходит из-за того, что люди чаще общаются и делятся информацией о местах отдыха, чем о предметах мебели.

**14.** Лингвометодические возможности использования корпусов текстов:

1. Автоматически выделять в двух языках группы слов определенного словоизменения или словообразования.

2. Искать и выделять слова с определенными грамматическими характеристиками.

3. Выделять грамматические модели некоторых понятий и их эквивалентов в переводном языке.

4. Выделять структурные модели словосочетаний исходного и переводного языков.

5. Проводить сопоставительный анализ двух языков на синтаксическом уровне.

6. Автоматически строить переводные конкордансы (словарь).

7. Верифицировать значения лексических единиц, уже зафиксированных в каких-либо словарях, особенно в том, что касается метафор и терминологических выражений.

8. Выделять новые устойчивые словосочетания и идиоматические выражения, которые целесообразно вводить в существующие словари и др.

9. Отбирать текстовые примеры для создания учебников и учебных пособий.

10. Создавать какие-либо словари по различным видам информации учебников и по всем материалам учебников.

11. Создавать двуязычные терминологические словари по различным дисциплинам.

12. Отбирать предложения и словосочетания определенных синтаксических структур.

13. Изучать стилистические особенности разных авторов.

14. Строить словари рифм.

15. Изучать ритмическую структуру строк, стихотворных текстов и др.

**Задание 3.**

Что такое корпус?

**1.** Создатели НКРЯ (Национального корпуса русского языка) приняли стратегию не включать переводы в основную часть корпуса, и это имеет свои основания. Основная цель НКРЯ - изучение и анализ русского языка на основе большого объема текстов на русском языке. Включение переводов может исказить данные и усложнить анализ, поскольку переводы могут содержать влияние других языков, а также различия в стилистике и грамматике.  
  
Однако, переводы могут быть полезными для изучения состояния языка в определенных случаях. Например:  
  
1. Сравнение переводов различных текстов на разные языки может помочь в изучении лексических и грамматических особенностей русского языка.  
  
2. Анализ переводов может помочь в изучении культурных особенностей и контекста, в котором используется определенное выражение или конструкция.  
  
3. Переводы могут быть полезны для сравнения стилей и манеры изложения текстов на разных языках.  
  
Таким образом, хотя НКРЯ не включает переводы в основную часть корпуса, они могут быть ценным дополнением для изучения языка и его использования в различных контекстах.

**2. Можно было бы создать корпус, в который вошли бы:**

1. Оригинальные тексты на английском и русском языках (литературные произведения, статьи, новости и т.д.).  
2. Переводы этих текстов с английского на русский и с русского на английский.  
3. Сравнительные анализы переводов различных переводчиков.  
4. Комментарии и аналитические статьи об особенностях перевода и лингвистических явлениях.  
  
Для того чтобы сделать такой корпус сбалансированным, нужно было бы учесть следующие критерии:  
  
1. Разнообразие жанров текстов: включение литературных произведений, научных статей, новостей и других типов текстов.  
2. Равное представительство текстов на английском и русском языках.  
3. Разнообразие авторов и переводчиков.  
4. Учет различных временных периодов и тематик текстов.  
5. Контроль за качеством переводов и соответствием оригиналу.

**3.**

Переводы глагола "to challenge" на русский язык:  
  
1. вызывать  
2. бросать вызов  
3. оспаривать  
4. предъявлять вызов  
5. ставить под сомнение  
6. вызов  
7. испытывать  
8. выставлять на испытание  
9. подвергать испытанию  
10. бросать вызов себе

**4.**

Целями авторов данного ресурса являются:  
  
1. Создание крупного корпуса текстов на русском языке, включающего разнообразные жанры и типы текстов.  
2. Проведение лингвистических исследований, анализа языковых явлений и особенностей русского языка.  
3. Поддержка различных проектов, связанных с анализом текстов и развитием компьютерной лингвистики.  
  
Авторы ГИКРЯ используют интернет для сбора текстов, так как он является богатым источником разнообразной информации на русском языке. Они проводят автоматизированный сбор текстов из различных веб-ресурсов, чтобы создать большой и разнообразный корпус для дальнейших исследований. Такой подход позволяет получить доступ к большому объему данных и анализировать их с помощью специализированных инструментов компьютерной лингвистики.

5.

Частотность слов ОПТИСИСТ и ПЕССИМИСТ значительно возросла по сравнению с прошлым веком.

Тест

Впишите слово или сочетание слов.

1 **Конкорданс** – это представленный в алфавитном порядке список всех

слов текста с указанием мест в тексте, в которых это слово встречается.

2 В современных лингвистических корпусах аналогом симфоний служит

формат выдачи результатов **конкорданс**

3 Словарь, составленный **на основе реального словоупотребления** является первым словарем

американского английского, созданным на основе реального

словоупотребления

4 Авторы **корпусных** грамматик описывают не норму, а реальное

состояние языка.

5 Противники корпусной лингвистики в оценке языковых явлений

предлагают опираться не на корпусные данные, а исключительно на

интроспекцию и языковую **интуицию**

6 Какие принципы составления лингвистического корпуса использованы

создателями НКРЯ?

**Значимые выводы можно делать только на основании материалов**

**корпуса достаточно большого объема.**

В национальном корпусе должны быть представлены несколько

разных языковых регистров.

Корпус создается для решения конкретной лингвистической задачи.

Пользователь имеет возможность получить информацию о каждом

входящем в корпус тексте.

7 В каких корпусах собраны тексты на русском языке?

Корпус Браун

Упсальский корпус

**СОСА**

Тюбингенский корпус

8 Какие утверждения описывают корпусную лингвистику в России?

**С момента своего появления развивается в тесном взаимодействии с**

**иностранными коллегами**.

Выработала стандартизированный формат сравнения разных

региональных вариантов языка.

Ввела в научный обиход понятие «национальный корпус языка».

Вырабатывает особую идеологию исследования языковых данных.

9 Какие утверждения описывают корпусную лингвистику в

Великобритании?

**С момента своего появления ставила целью исследование**

**современного языка.**

Впервые продемонстрировала методы машинной разметки текстов.

Разрабатывается активными последователями Н. Хомского.

Провозглашает первостепенную значимость интроспекции для

изучения грамматических конструкций.

10 Какие области исследования являются особенно привлекательными

для корпусных лингвистов?

Изучение регистров языка.

**Описание и установление нормы**.

Изучение диахронических языковых изменений.

Изучение контекстуальной вариативности лексических значений.

11 Какие корпуса дают материал для изучения американского

английского языка?

Корпус Браун

**BNC**

СОСА

LOB Corpus

Выберите единственный правильный

вариант ответа.

12 Национальный корпус НЕ

включает поэтических текстов.

отражает диалектные особенности говоров.

отражает состояние языка страны в конкретные временные периоды.

**отражает государственную идеологию страны.**

13 Sample corpus

должен состоять из образцово грамотных текстов.

**отражает состояние языка данной местности в данный период.**

не может включать целые тексты и составлен исключительно из

фрагментов.

не может быть создан на материале русского языка.

14 На материале сравнимых корпусов обычно изучается

словообразование.

морфология.

синтаксис.

**вариативность.**

15 Авторы некоторых корпусов отказываются включать в них

драматические произведения, поскольку

такие произведения сложно переводить в электронный формат.

**речь персонажей является имитацией устной речи и не может дать**

**представления о реальном языке диалога.**

объем корпуса значительно увеличивается за счет введения в них

драмы.

драматические авторы редко вносят заметный вклад в развитие

литературного языка.

**Разметка корпусов**

1.Разметка корпуса должна быть теоретически нейтральной, чтобы обеспечить объективность и точность анализа текстов. Нейтральная разметка позволяет исследователям использовать данные корпуса для различных целей без искажения результатов и предвзятости.  
  
2. Грамматический словарь А. Зализняка содержит информацию о морфологических характеристиках слов русского языка, таких как основа, окончания, грамматические категории и т.д. Эти данные могут быть использованы для автоматической морфологической разметки текстов путем сопоставления слов из текста с данными из словаря и определения их морфологических характеристик.  
  
3. Примеры типов омонимии, которые не всегда могут быть автоматически сняты при разметке:  
- Грамматическая омонимия: "лук" (оружие) и "лук" (овощ);  
- Лексическая омонимия: "банка" (посуда) и "банка" (организация).  
  
4. Высказывание авторов о том, что русская морфология представленная в НКРЯ более "либеральная" и широкая, чем в нормативных грамматиках, может быть объяснено тем, что корпус учитывает разнообразные варианты использования слов в реальных текстах, включая различные формы, устаревшие или диалектные слова, что может отличаться от строгих норм грамматики.  
  
5. Пример поискового запроса, сочетающего грамматические и семантические категории в НКРЯ: "существительное в родительном падеже, обозначающее предметы для еды". Такой запрос поможет пользователю отсеять лишние контексты и сузить результаты поиска до конкретного семантического поля.

**Поиск информации в корпусе**

1.Список низкочастотных коллокаций может помочь выявить, какие словосочетания могут быть сложными для усвоения из-за их редкости и нестандартности.  
  
2. Для создания подкорпуса в НКРЯ, соответствующего исследовательским задачам, можно учитывать следующие параметры текстов:  
- Жанр текстов (научные статьи, художественная литература и т.д.);  
- Временной период написания текстов;  
- Авторство текстов;  
- Тематика текстов;  
- Региональные особенности языка.  
  
3. Для доказательства того, что Андрей Болконский мог видеть единорогов, можно провести поиск в НКРЯ по упоминаниям этого персонажа из произведения Льва Толстого "Война и мир". При нахождении упоминаний описаний сцен или диалогов, где говорится о видении единорогов Андреем Болконским, можно подтвердить данное утверждение.  
  
4. Для выяснения времени вхождения лексемы в язык в НКРЯ можно использовать функцию "Этимологический анализ слова". Эта функция позволяет узнать происхождение слова, его историю и первоначальное значение, что поможет определить, когда и как лексема вошла в русский язык.  
  
5. Для доказательства того, что проект "Русская грамматика" является дескриптивной, а не нормативной грамматикой, можно обратиться к описанию языковых явлений и правил в проекте. Если грамматика описывает фактическое употребление языка на основе данных из корпусов и анализа текстов, то это скорее всего дескриптивная грамматика. В случае нормативной грамматики акцент делается на правилах и нормах языка, которые должны соблюдаться.

**Тест**

Впишите слово.

1 С помощью **статистических** методов исследователь выясняет, случайное или закономерное

явление он наблюдает.

2 Результаты корпусных исследований обычно представляются в виде **относительных** значений,

то есть количества слов на миллион.

3 Слова, которые встречаются в данном корпусе или тексте чаще, чем они обычно

встречаются в языке, называются **ключевыми** словами.

4 Выявить такие частотные слова можно только используя два корпуса: тот, на материале

которого проводится исследование, и **референтным** корпус.

5 Современные методики английского языка предлагают ученику запоминать не только

отдельные лексемы, но и неслучайные сочетания слов, то есть **коллокации**

Выберите один или несколько верных вариантов

ответа.

Будьте внимательны! В некоторых случаях правильными могут оказаться все предложенные

варианты.

6 В ходе corpus-based исследования

в корпусе подбираются примеры, подтверждающие гипотезу автора.

**автор формулирует гипотезу только после анализа результатов, полученных в корпусе.**

автор проверяет и дополняет свою гипотезу.

автор действует методом индукции.

7 Corpus-driven подход к изучению языка

предлагает отбросить исключения и сосредоточиться на анализе правил.

**выстраивает гипотезу, базируясь на реальном словоупотреблении.**

может быть назван дедуктивным.

может отвергать значимость анализа лексем и основываться на изучении словоформ.

8 Материалом corpus-driven исследования часто становится

группа контекстов, демонстрирующих языковую норму.

**языковое явление, которое можно назвать ключевым для изучаемой группы текстов**

**референтный корпус.**

языковое явление, которое демонстрируют тексты художественной литературы.

9 Результаты corpus-based и corpus-driven исследований

**не позволяют изменить и расширить существующие лингвистические теории.**

**обычно не согласуются с выводами, сделанными традиционными лингвистами.**

подтверждают мысль о невозможности анализа языковых явлений вне конкретного

**регистра.**

всегда основаны на анализе частотности изучаемого явления.

10 Лингвистический корпус может помочь преподавателю языка

определить содержание учебного курса.

установить порядок изучения материала.

**определить количество заданий, необходимое для освоения материала.**

**найти иллюстрации для изучаемых правил.**

Выберите единственный правильный вариант ответа.

11 Величина частотности конструкций

может быть оценена с помощью интроспекции.

определяется интуитивно.

**выясняется с помощью статистических методов.**

не значима в процессе обучения языку.

12 Учебные пособия, базирующиеся на корпусных данных,

не используются в младшей и средней школе.

часто содержат сведения о вариативности.

не могут быть использованы при обучении родному языку.

**предполагают знакомство ученика с методами работы в лингвистическом корпусе.**

13 Кэмбриджский корпус английского языка

включает тексты людей, плохо пишущих по-английски.

не имеет коммерческой ценности.

используется всеми британскими издательствами.

**находится в открытом доступе.**

14 Национальный корпус русского языка

будучи общедоступным ресурсом, не может применяться в преподавании.

**является ценным источником примеров для составителей учебных материалов**.

стал основой для нескольких серий школьных учебников.

может использоваться в педагогической практике только учителями, имеющими

серьезную лингвистическую подготовку.

15 Преподаватель иностранного языка, пользующийся корпусными методиками, НЕ

структурирует время, выделяемое на изучение материала, в зависимости от частотности

явления.

предлагает ученикам сведения об отличиях устной речи от письменной.

предлагает для запоминания коллокации.

**ограничивается анализом нормативных языковых единиц.**