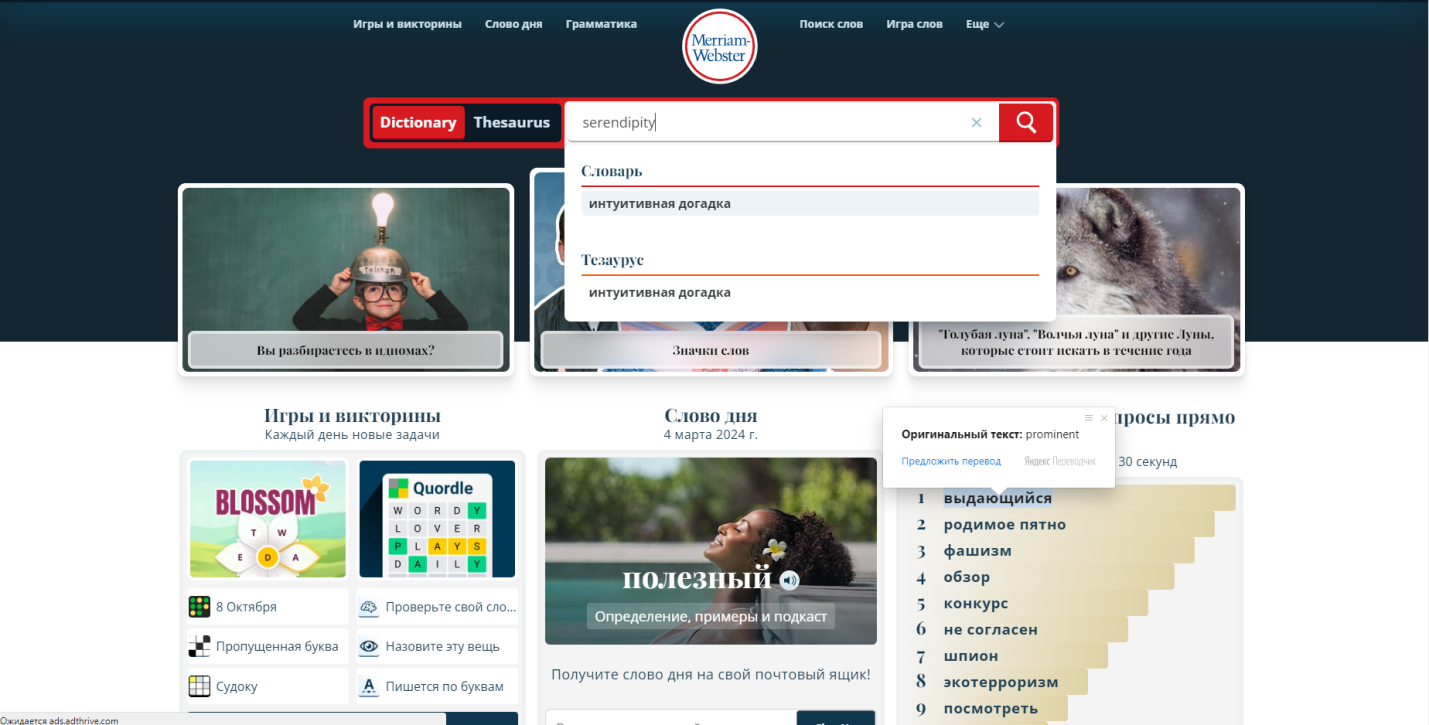
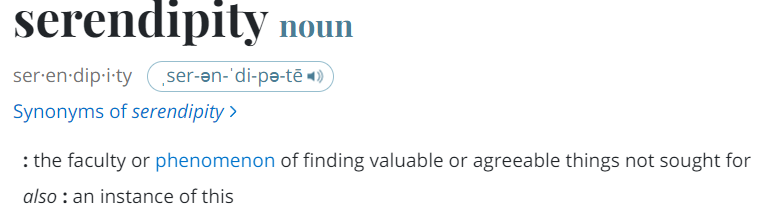
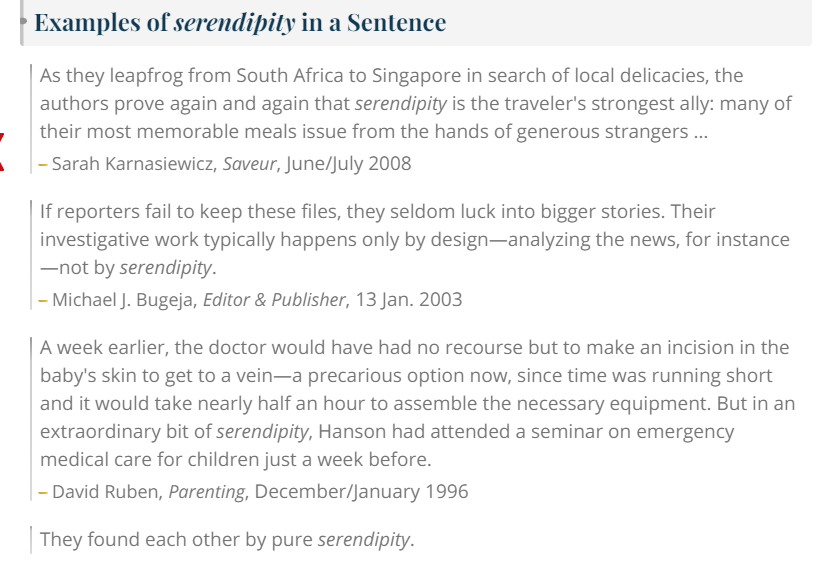
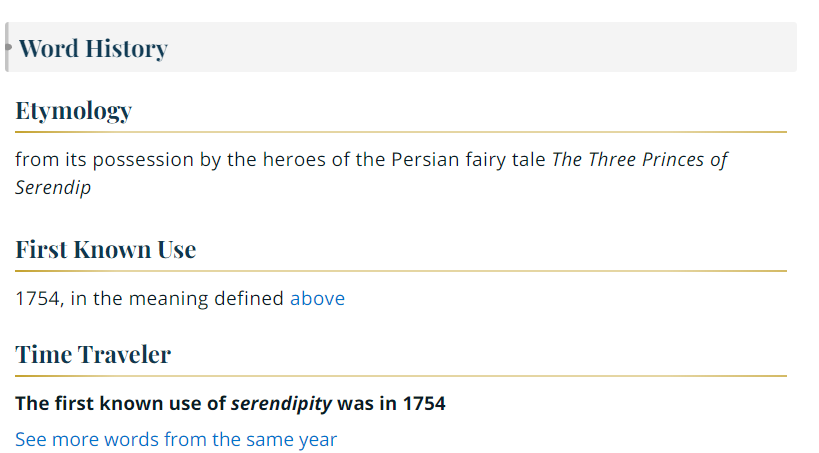
Уровень 1.

Задание 1.

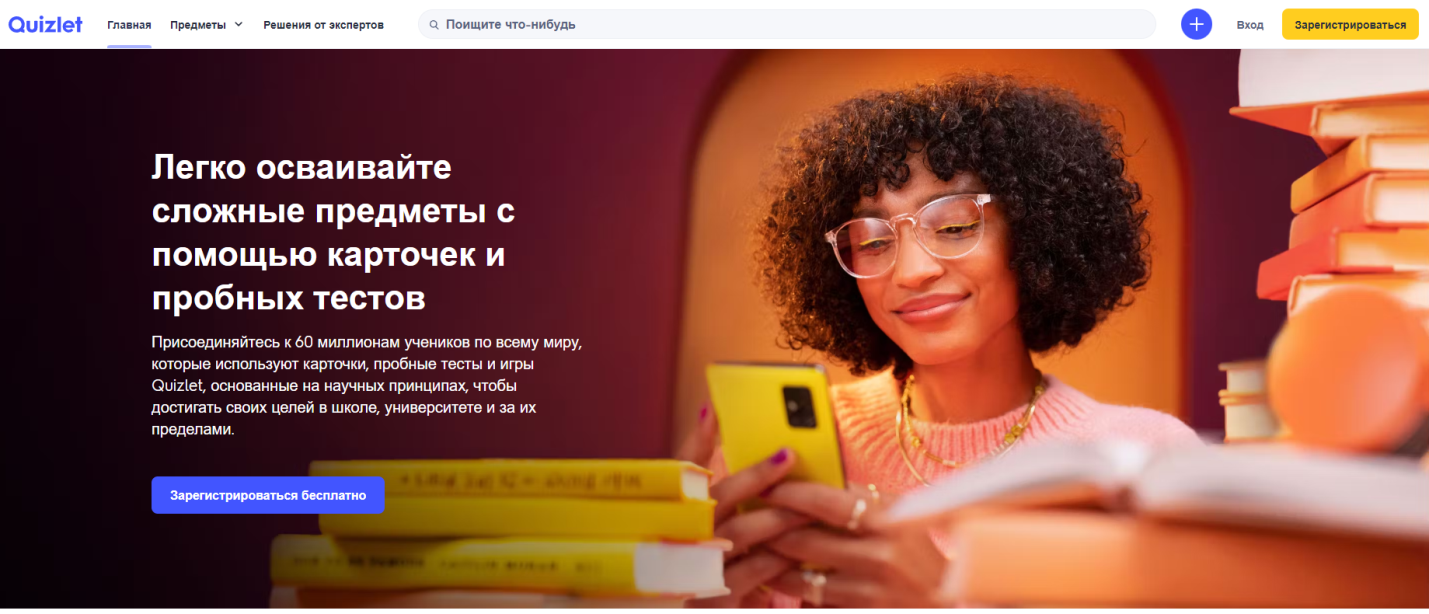
Изучив веб-сайт Merriam-Webster ([https://www.merriam-webster.com/](https://vk.com/away.php?utf=1&to=https%3A%2F%2Fwww.merriam-webster.com%2F)), можно отметить следующие возможности и информацию, представленную на нем:  
  
1. Информация: Merriam-Webster - это онлайн-словарь и энциклопедия, предоставляющая определения слов, их происхождение, синонимы, антонимы, примеры использования и многое другое.  
  
2. Использование: Ресурс можно использовать для проверки правописания и значений слов, изучения новых слов, поиска синонимов и антонимов, а также для обучения английскому языку.  
  
3. Запросы: Для поиска определения слова или фразы на сайте Merriam-Webster, просто введите его в поле поиска на главной странице.  
  
4. Преимущества: Merriam-Webster предлагает широкий выбор слов и высокое качество определений. Сайт легко доступен и удобен в использовании.  
  
5. Недостатки: Некоторые определения могут быть слишком сложными для начинающих изучать английский язык.  
  
6. Похожие ресурсы: Cambridge Dictionary ([https://dictionary.cambridge.org/](https://vk.com/away.php?utf=1&to=https%3A%2F%2Fdictionary.cambridge.org%2F)), Oxford English Dictionary ([https://www.oed.com/](https://vk.com/away.php?utf=1&to=https%3A%2F%2Fwww.oed.com%2F)).  
  
7. Исследования: За последние три года на базе Merriam-Webster проводились исследования по развитию лексической базы данных, улучшению интерфейса для пользователей и добавлению новых функций.  
  
Пример работы:  
- Поисковый запрос: "serendipity"

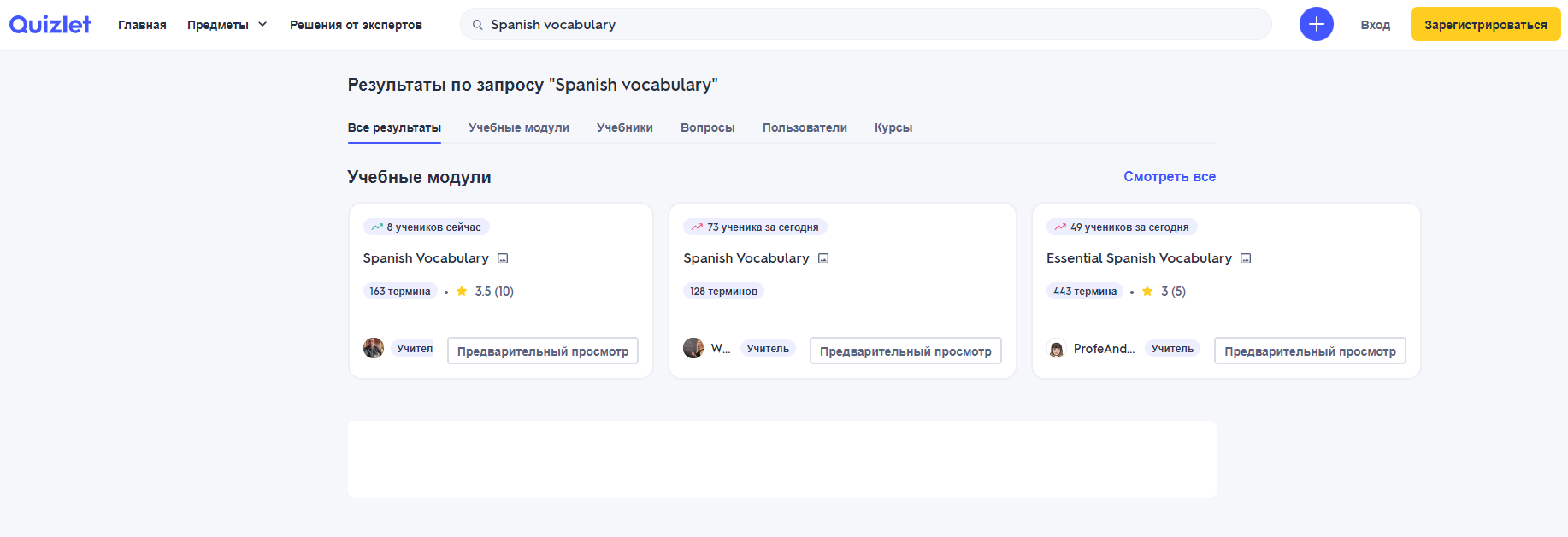
  
- Результаты: Определение слова "serendipity", происхождение, примеры использования.



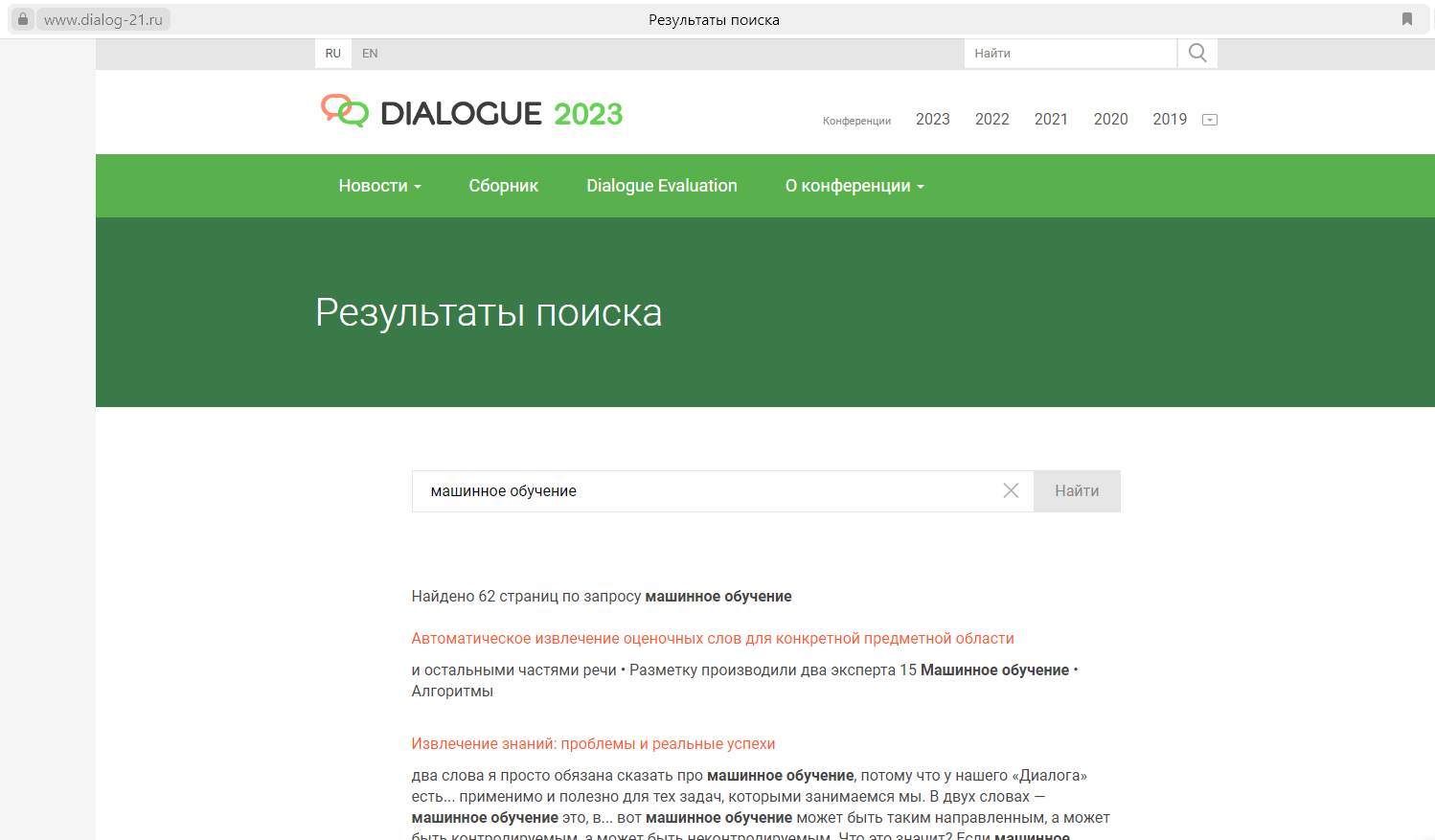
   
- Анализ: Поиск выдал определение слова "serendipity" как "the faculty or phenomenon of finding valuable or agreeable things not sought for". Это позволяет пользователям узнать значение и использование данного слова в контексте.

Изучив веб-сайт Quizlet ([https://quizlet.com/](https://vk.com/away.php?utf=1&to=https%3A%2F%2Fquizlet.com%2F)), можно отметить следующие возможности и информацию, представленную на нем:  
  
1. Информация: Quizlet - это онлайн-платформа для обучения и запоминания информации с помощью карточек, тестов и других учебных материалов. Пользователи могут создавать свои собственные наборы карточек или использовать уже существующие.  
  
2. Использование: Ресурс можно использовать для изучения новых тем, подготовки к тестам и экзаменам, запоминания информации, а также для обмена учебными материалами с другими пользователями.  
  
3. Запросы: Для поиска конкретного набора карточек или теста на сайте Quizlet, можно воспользоваться поиском по ключевым словам или теме.  
  
4. Преимущества: Quizlet предлагает широкий выбор учебных материалов, интерактивные способы обучения, возможность создания собственных наборов карточек и тестов.  
  
5. Недостатки: Некоторые учебные материалы могут быть непроверенными и содержать ошибки.  
  
6. Похожие ресурсы: Anki ([https://apps.ankiweb.net/](https://vk.com/away.php?utf=1&to=https%3A%2F%2Fapps.ankiweb.net%2F)), Memrise ([https://www.memrise.com/](https://vk.com/away.php?utf=1&to=https%3A%2F%2Fwww.memrise.com%2F)).  
  
7. Исследования: За последние три года на базе Quizlet проводились исследования по эффективности методов обучения через карточки, развитию функционала платформы для улучшения процесса запоминания информации.  
  
Пример работы:  
- Поисковый запрос: "Spanish vocabulary"

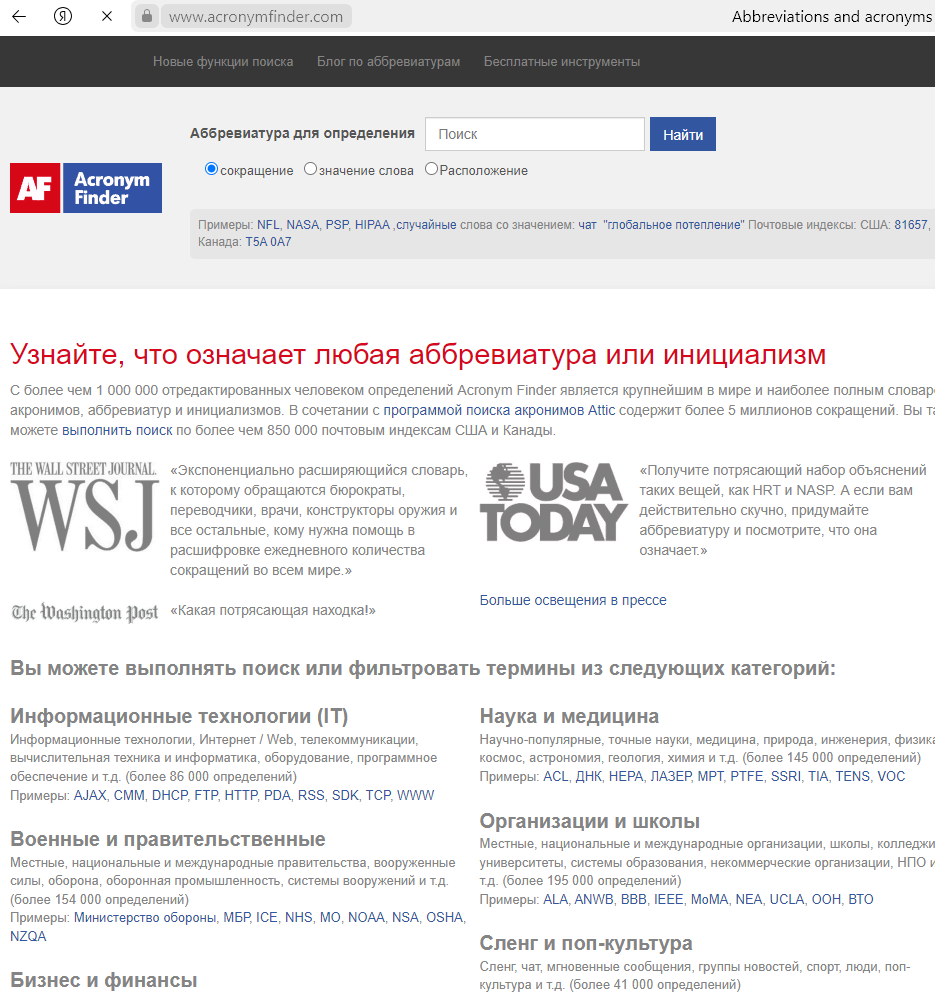
  
- Результаты: Найдены несколько наборов карточек с испанской лексикой, включая основные слова и фразы.

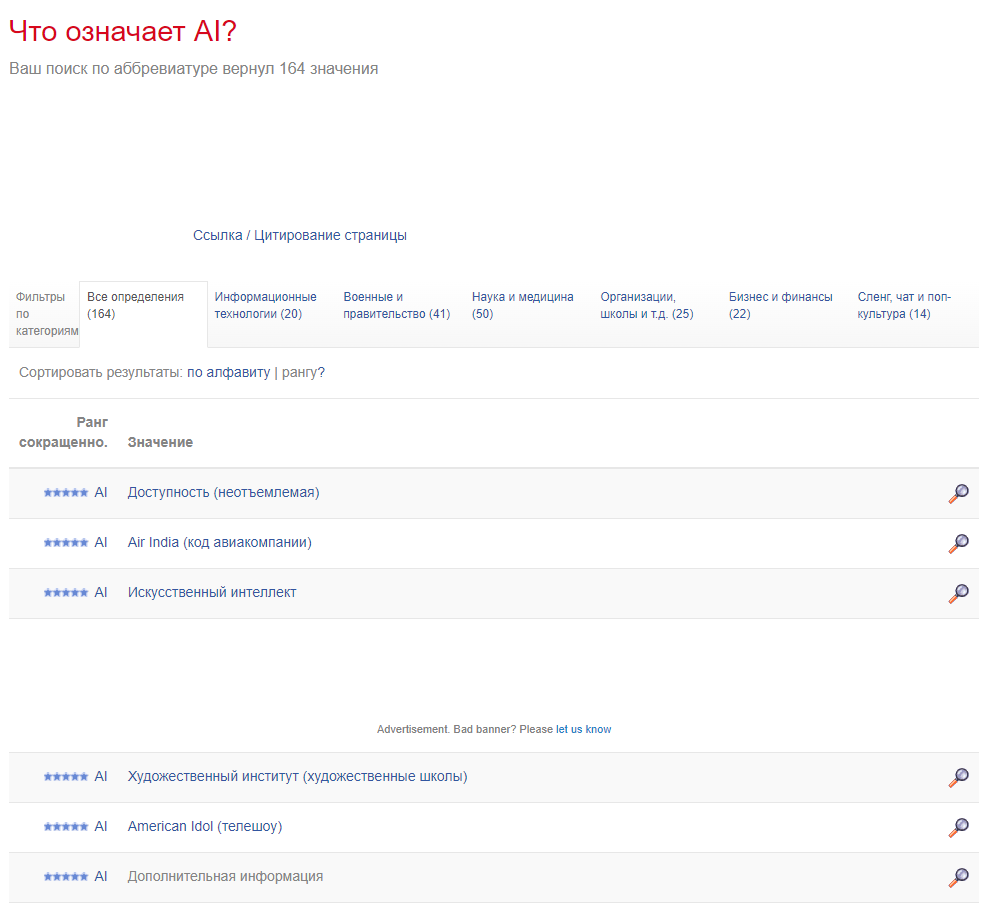
  
- Анализ: Поиск позволил найти учебные материалы по испанскому языку, которые могут быть полезны для изучения новой лексики. Пользователи могут использовать эти карточки для тренировки своих знаний и запоминания слов и выражений на испанском языке.

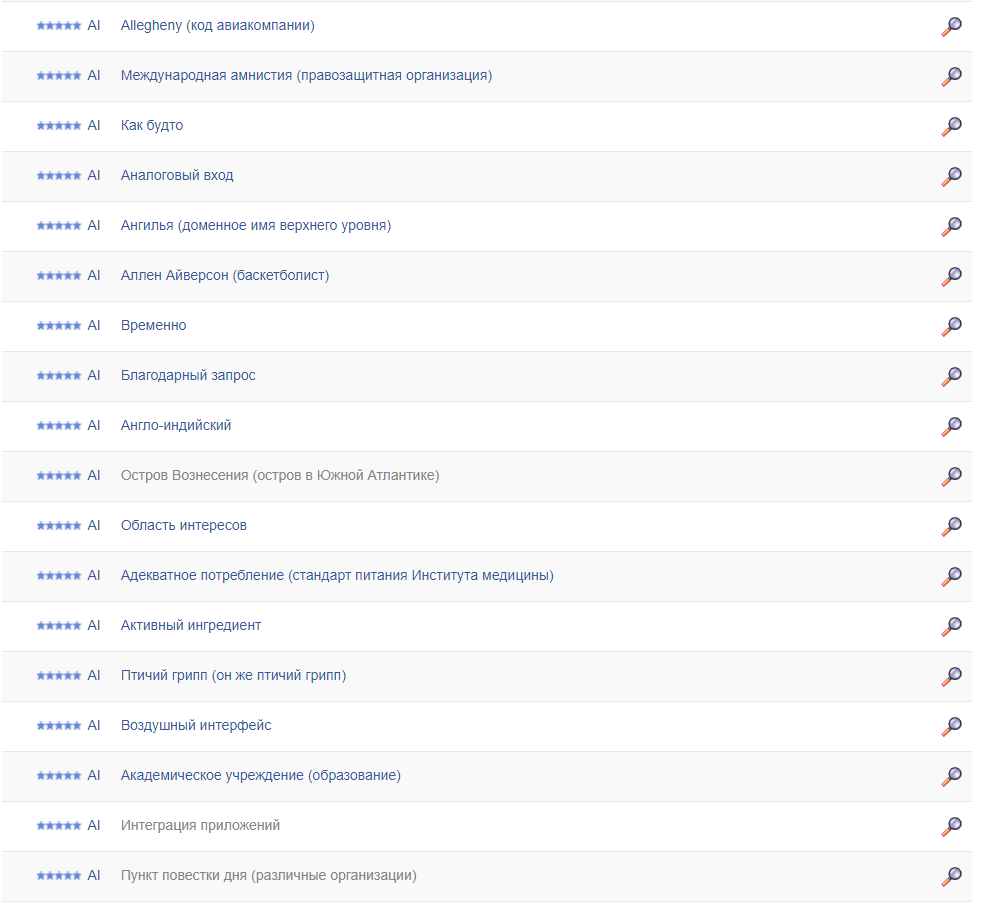
Ресурс [http://www.dialog-21.ru/](https://vk.com/away.php?utf=1&to=http%3A%2F%2Fwww.dialog-21.ru%2F) представляет собой официальный сайт конференции "Диалог", посвященной компьютерной лингвистике. На сайте представлена информация о конференции, программа мероприятий, докладчиках, публикациях и других актуальных событиях в области компьютерной лингвистики.  
  
Этот ресурс можно использовать для следующих целей:  
- Получения информации о последних тенденциях и исследованиях в области компьютерной лингвистики  
- Участия в конференции и презентации своих научных разработок  
- Поиска партнеров для совместных исследований  
- Обмена опытом с коллегами из разных стран  
  
Для производства запросов на сайте [http://www.dialog-21.ru/](https://vk.com/away.php?utf=1&to=http%3A%2F%2Fwww.dialog-21.ru%2F) необходимо использовать стандартные функции поиска, которые обычно представлены на главной странице или в специальном разделе "Поиск".  
  
Преимущества данного ресурса:  
- Актуальная информация о конференции и связанных с ней событиях  
- Возможность участия в международном научном сообществе  
- Доступ к публикациям и материалам конференции  
  
Недостатки:  
- Ограниченный функционал для поиска информации  
- Не всегда доступны все материалы и доклады конференции  
  
За последние 3 года на базе этого ресурса проводились многочисленные исследования в области компьютерной лингвистики, касающиеся различных аспектов обработки естественного языка, машинного перевода, анализа текстов и других тем.  
  
Пример работы:  
Предположим, что я хочу найти всю возможную информацию о машинном переводе. Мой поисковый запрос может выглядеть следующим образом: "Машинный перевод". Результаты поиска покажут список статей, которые можно изучить для дальнейшего анализа и использования в своих исследованиях.



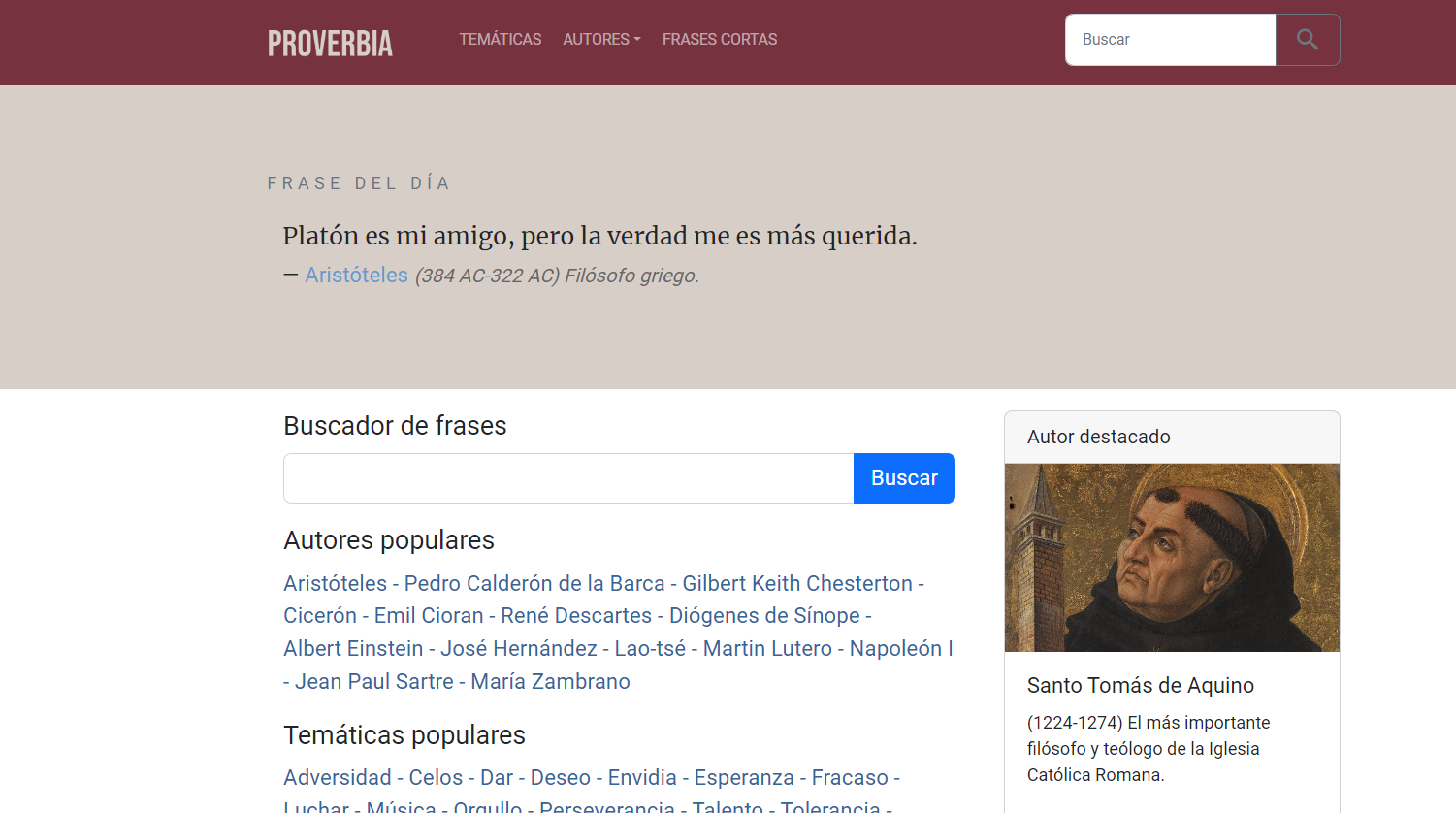
Ресурс [http://acronymfinder.com/](https://vk.com/away.php?utf=1&to=http%3A%2F%2Facronymfinder.com%2F) представляет собой онлайн словарь аббревиатур и акронимов, который помогает пользователям найти расшифровку значений различных сокращений. На сайте представлена обширная база данных, включающая в себя аббревиатуры из различных областей, таких как медицина, технологии, бизнес, образование и другие.  
  
Основные возможности ресурса:  
- Поиск расшифровок аббревиатур и акронимов  
- Просмотр различных вариантов значений для одной аббревиатуры  
- Обновление базы данных с новыми сокращениями  
  
Этот ресурс можно использовать для следующих целей:  
- Поиска значения неизвестных аббревиатур  
- Расшифровки сокращений в тексте или документах  
- Поиска синонимов или альтернативных интерпретаций аббревиатур  
  
Для производства запросов на сайте [http://acronymfinder.com/](https://vk.com/away.php?utf=1&to=http%3A%2F%2Facronymfinder.com%2F) необходимо использовать стандартную функцию поиска, которая обычно представлена на главной странице. Вводите аббревиатуру или акроним, для которого хотите получить расшифровку, и нажимайте кнопку "Search".  
  
Преимущества данного ресурса:  
- Большая база данных сокращений из различных областей  
- Возможность быстро найти значение неизвестных аббревиатур  
- Простой и удобный интерфейс для поиска  
  
Недостатки:  
- Не всегда представлена полная информация об аббревиатурах  
- Могут быть ограничения в доступе к некоторым данным  
  
За последние 3 года на базе этого ресурса проводились исследования по анализу частоты использования различных аббревиатур в различных областях, а также по разработке методов автоматического распознавания и классификации сокращений.  
  
Пример работы:  
Предположим, что я хочу узнать значение аббревиатуры "AI". Мой поисковый запрос будет выглядеть следующим образом: "AI". Результаты покажут различные значения этой аббревиатуры, например, "Artificial Intelligence", "Aviation Institute" и другие. Я могу выбрать нужное значение и получить расшифровку аббревиатуры.







Ресурс [www.proverbia.net](https://vk.com/away.php?utf=1&to=http%3A%2F%2Fwww.proverbia.net) является онлайн платформой, где пользователи могут найти цитаты, пословицы и поговорки на различные темы. Сайт предлагает обширную базу данных с цитатами из разных культур и эпох, которые могут быть использованы для вдохновения, обучения, цитирования и других целей.  
  
Основные возможности ресурса:  
- Поиск цитат по ключевым словам, авторам, темам  
- Просмотр случайных цитат  
- Сохранение и поделиться цитатами  
- Создание списка любимых цитат  
  
Этот ресурс можно использовать для следующих целей:  
- Нахождения мудрых высказываний для вдохновения и мотивации  
- Изучения культурных особенностей через пословицы и поговорки  
- Добавления цитат в презентации, эссе, статьи и т.д.  
  
Для производства запросов на сайте [www.proverbia.net](https://vk.com/away.php?utf=1&to=http%3A%2F%2Fwww.proverbia.net) необходимо использовать функцию поиска, которая обычно представлена на главной странице. Вводите ключевое слово, автора или тему цитаты, которую хотите найти, и нажимайте кнопку "Search".  
  
Преимущества данного ресурса:  
- Разнообразие цитат из разных культур и эпох  
- Возможность быстро найти нужную цитату по ключевым словам  
- Удобный интерфейс для просмотра и сохранения цитат  
  
Недостатки:  
- Некоторые цитаты могут быть неполными или содержать ошибки  
- Ограниченный функционал для пользователей  
  
За последние 3 года на базе этого ресурса проводились исследования по анализу популярности различных цитат, выявлению тенденций в использовании пословиц и поговорок в различных контекстах, а также разработке методов автоматического сбора и анализа цитат.  
  
Пример работы:  
Предположим, что я хочу найти пословицу о дружбе. Мой поисковый запрос будет выглядеть следующим образом: "friendship". Результаты покажут различные пословицы о дружбе, например, "A friend in need is a friend indeed". Я могу выбрать нужную пословицу и использовать ее для своих целей.



Задание 2

1.

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

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

2.

|  |  |  |  |
| --- | --- | --- | --- |
| № примера | НКРЯ | COCA | BNC |
| 1 | А русский сделал? Мне учительница сказала, что она тебе двойку в четверти поставит: ты делаешь по три ошибки в слове(Елена Павлова. Вместе мы эту пропасть одолеем! // «Даша», 2004 ) | He called me a **Russian** agent.(2019 SPOK [Fox\_Carlson](https://www.english-corpora.org/Coca/x4.asp?t=5001651&ID=856400305)) | Like **Russian** president Boris Yeltsin who still stands accused of continuing the cover-up over the execution while([W\_newsp\_tabloid](https://www.english-corpora.org/bnc/x4.asp?t=1208&ID=33822708" \t "x4)) |
| 2 | Типично французский вопрос: " Бенедикт, к тебе приезжал русский друг из Москвы, а на каком языке ты с ним общалась? "(Запись LiveJournal (2004)) | allegedly shared with Sergey Lavrov, the foreign minister, Sergey Kislyak, the Russian ambassador to the United States. Israel would have a deep and abiding concern that([CNN\_Newsroom](https://www.english-corpora.org/Coca/x4.asp?t=4172700&ID=833700106)2017 SPOK) | ven so, that doesn't stop him being introduced by **Russian** coach Anton Pamchenko (Roy Dotrice) to bitchy ice-queen Kate Moseley (Moira Kelly(([W\_newsp\_tabloid](https://www.english-corpora.org/bnc/x4.asp?t=1208&ID=33822708" \t "x4)) |
| 3 | В Эстонии — эстонский для русских, русский для эстонцев.( Запись LiveJournal с комментариями) | According to **Russian** consultancy Zest, this change has seen the share of Russian-owned aircraft that are also([ainonline.com](https://www.english-corpora.org/Coca/x4.asp?t=5119405&ID=1335900609)2019 SPOK) | The terrorists subjected grandmother, mother and child to **Russian** roulette.' It is thought the bomb was aimed at a member of the([W\_newsp\_tabloid](https://www.english-corpora.org/bnc/x4.asp?t=1208&ID=33822708" \t "x4)) |

3.

|  |  |  |  |
| --- | --- | --- | --- |
| № примера | НКРЯ | COCA | BNC |
| 1 | Сущ | Прил | Прил |
| 2 | Прил | Прил | Прил |
| 3 | Сущ, прил | Прил | Прил |

 4.

|  |  |  |  |
| --- | --- | --- | --- |
| № примера | НКРЯ | COCA | BNC |
| 1 | Школьного предмета | Относящийся к Росси́и | Относящийся к Росси́и |
| 2 | Относящийся к Росси́и | Относящийся к Росси́и | Относящийся к Росси́и |
| 3 | Человек, относящийся к России | Относящийся к Росси́и | Относящийся к Росси́и |

5.

К каким выводам вы пришли при сравнении морфологической и лексической характеристики одного и того же слова, включенного в разные корпусы?

При сравнении морфологических и лексических характеристик одного и того же слова, включенного в разные корпусы, можно прийти к следующим выводам:  
  
1. Морфологические характеристики (такие как часть речи, род, число, падеж и другие) остаются постоянными для данного слова в разных корпусах. Это связано с тем, что морфологические характеристики определяются внутренней структурой слова и не зависят от контекста, в котором оно используется.  
  
2. Лексические характеристики (такие как значение, семантика, употребление в определенных контекстах) могут различаться в разных корпусах, так как они зависят от конкретного контекста, в котором слово употребляется. Например, слово "русский язык" может использоваться в разных контекстах (научном, образовательном, повседневном и т. д.), что может влиять на его лексические характеристики.  
  
Таким образом, сравнение морфологических и лексических характеристик слова в разных корпусах позволяет выявить стабильные морфологические особенности и изменчивые лексические значения данного слова в различных контекстах использования.

6.

Рассмотренные корпусы текстов могут быть полезны в различных лингвистических исследованиях. Вот несколько способов, как можно использовать корпусы в лингвистическом исследовании:  
  
1. Лексикография: Корпусы могут использоваться для составления словарей, тезаурусов и других лексикографических ресурсов. Анализ частотности употребления слов, их семантических связей и контекстов помогает лексикографам создавать более точные и информативные словари.  
  
2. Семантические исследования: Корпусы позволяют изучать значения слов, их употребление в различных контекстах, семантические поля и изменения значений со временем. Это помогает лингвистам понять семантическую структуру языка.  
  
3. Синтаксические исследования: Корпусы текстов предоставляют данные для анализа синтаксических структур, построения деревьев зависимостей, изучения синтаксических конструкций и грамматических особенностей языка.  
  
4. Исследования стилистики и стилометрии: Корпусы могут использоваться для анализа стилей авторов, выявления стилистических особенностей текстов, а также для автоматической стилометрии (определения авторства текстов).  
  
5. Исследования языковых изменений: Корпусы позволяют отслеживать изменения в языке со временем, изучать эволюцию лексики, грамматики и семантики.  
  
6. Машинный перевод и обработка естественного языка: Корпусы используются для обучения моделей машинного перевода, анализа тональности текстов, распознавания именованных сущностей и других задач обработки естественного языка.  
  
Это лишь некоторые примеры того, как можно использовать корпусы текстов в лингвистических исследованиях. Важно уметь адаптировать методы работы с корпусами под конкретные цели и задачи исследования.

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) Национальный корпус русского литературного языка (НКРЯ) является крупным корпусом текстов на русском языке, который содержит разнообразные литературные произведения, от древнерусских до современных. Количество словоупотреблений в НКРЯ огромное, составляя миллионы слов.  
  
Вид корпуса НКРЯ можно характеризовать как:  
  
1. Литературный: в нем собраны произведения русской литературы разных эпох и жанров, что позволяет изучать языковые особенности литературного стиля.  
  
2. Исторический: в корпусе представлены тексты различных исторических периодов, что позволяет исследовать изменения в русском языке на протяжении времени.  
  
3. Репрезентативный: НКРЯ содержит широкий спектр текстов, что делает его репрезентативным для изучения различных языковых явлений.  
  
4. Электронный: корпус доступен в электронном виде через интернет, что облегчает доступ к данным и проведение лингвистических исследований.

Количество словоупотреблений в Национальном корпусе русского литературного языка (НКРЯ) огромное и составляет миллионы. Точное число словоупотреблений в корпусе может меняться в зависимости от объема текстов, которые включены в его состав, но общее количество текстов и слов в НКРЯ делает его одним из крупнейших корпусов русского языка для лингвистических исследований.  
  
Таким образом, Национальный корпус русского литературного языка представляет собой ценный ресурс для лингвистических исследований благодаря своему обширному объему текстов и разнообразию жанров и эпох.

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) 1. Sketch Engine ([https://www.sketchengine.eu/](https://vk.com/away.php?utf=1&to=https%3A%2F%2Fwww.sketchengine.eu%2F)) – это онлайн-платформа для работы с корпусами текстов, предоставляющая инструменты для анализа лексики, грамматики и стилистики текстов на различных языках. Sketch Engine позволяет проводить разнообразные лингвистические исследования, включая изучение коллокаций, конкордансов, частотности слов и многое другое.  
  
2. Corpus of Contemporary American English (COCA) ([https://www.english-corpora.org/coca/](https://vk.com/away.php?utf=1&to=https%3A%2F%2Fwww.english-corpora.org%2Fcoca%2F)) – это крупный корпус английских текстов современного американского английского языка. COCA содержит миллионы словосочетаний и предложений, позволяя исследователям изучать употребление слов и выражений в различных контекстах.  
  
3. Linguistic Data Consortium (LDC) ([https://www.ldc.upenn.edu/](https://vk.com/away.php?utf=1&to=https%3A%2F%2Fwww.ldc.upenn.edu%2F)) – это организация, специализирующаяся на сборе, разработке и распространении лингвистических данных и корпусов текстов на различных языках. LDC предлагает доступ к широкому спектру корпусов текстов для исследований в области корпусной лингвистики.  
  
4. National Corpus of Polish (NKJP) ([https://nkjp.pl/](https://vk.com/away.php?utf=1&to=https%3A%2F%2Fnkjp.pl%2F)) – это крупный корпус польских текстов, который содержит разнообразные тексты на польском языке, начиная с XIX века. NKJP предоставляет возможность исследования языковых явлений и особенностей польского языка с использованием большого объема текстовых данных.  
  
5. British National Corpus (BNC) ([https://www.natcorp.ox.ac.uk/](https://vk.com/away.php?utf=1&to=https%3A%2F%2Fwww.natcorp.ox.ac.uk%2F)) – это корпус британского английского языка, который содержит разнообразные тексты на английском языке, собранные в период с 1980 по 1993 год. BNC предоставляет исследователям доступ к большому объему текстов для изучения различных языковых явлений и особенностей британского английского.

Уровень 2

|  |  |
| --- | --- |
| **Корпусная лингвистика** | |
| В чем сущность эмпирической методологии в лингвистическом анализе? | Эмпирическая методология в лингвистическом анализе основана на использовании наблюдений, экспериментов и сборе данных для проверки гипотез и теорий о языке. Она предполагает сбор конкретных фактов о языке, их анализ и интерпретацию на основе эмпирических данных.  Сущность эмпирической методологии в лингвистическом анализе заключается в том, что она позволяет лингвистам проверять свои гипотезы и теории на основе конкретных данных, полученных из наблюдений и экспериментов. Это помогает уточнять и развивать научные представления о языке и его структуре, а также делать более обоснованные выводы о языковых явлениях. |
| Дайте определение понятию коллокация | Коллокация (collocation) — **это устойчивое словосочетание, которое представляет собой целостную единицу как с точки зрения семантики, так и в плане синтаксиса.** |
| Сформулируйте закон Ципфа. | **Закон Ципфа** — это эмпирическая закономерность распределения частотности слов естественного языка.  Она заключается в следующем: если все слова языка (или просто достаточно длинного текста) упорядочить по убыванию частотности их использования, то частотность n-го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру n (так называемому рангу этого слова). |
| Дайте определениепонятию дискурсивной просодии | Дискурсивная просодия относится к изучению того, как интонация, ритм, темп и другие аспекты звуковой структуры речи используются для передачи смысла и организации дискурса. Это включает в себя исследование того, как изменения в произношении слов или фраз могут влиять на интерпретацию сообщения, выражение эмоций, акцентирование ключевых моментов и управление вниманием слушателей. Дискурсивная просодия является важным аспектом коммуникации и может помочь понять, каким образом звуковые характеристики речи влияют на ее восприятие и понимание. |
| Дайте определение понятию семантической просодии | Одним из важных терминов корпусной лингвистики является понятие семантической просодии (semantic prosody), определяемой как оценочное отношение говорящего к содержанию высказывания, обозначение информации как позитивно или негативно окрашенной. |
| В чём сущность гипотизы общего ядра | Гипотеза общего ядра (Common Core Hypothesis) в компьютерной лингвистике предполагает, что существует некоторое общее ядро лексических единиц или концептов, которые являются универсальными для всех языков. Это общее ядро может быть использовано для создания универсальных языковых ресурсов, моделей и инструментов, которые могут быть применены к различным языкам.  Идея заключается в том, что существует определенное количество базовых концептов или слов, которые присутствуют практически во всех языках, и что эти концепты могут быть использованы как основа для создания универсальных языковых моделей. Это позволяет упростить разработку и сравнение языковых систем, а также облегчает перевод и анализ текстов на разных языках.  Гипотеза общего ядра активно исследуется в области компьютерной лингвистики и искусственного интеллекта, где разрабатываются методы и технологии для создания универсальных языковых моделей на основе этой концепции. |
| Что такое дерево зависимостей? | **Деревья** **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? | Реферат, резюме, конспект, 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. |
| Определение понятия 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? | Алфавитный перечень всех слов какого-либо текста с указанием контекстов их употребления; тип словаря, представленный в таком виде |
| Что такое коркондансер? | **Конкордансер** — это компьютерная программа для автоматического составления конкорданса, то есть списка примеров употребления слова в минимальном контексте, или перечня всех контекстов, в которых слово или словосочетание встречается в исследуемом тексте. |
| 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. Что такое закон Ципфа?  
**a) Закон о распределении частот слов в тексте**  
b) Математическая формула для вычисления частоты слов  
c) Теория о структуре предложений  
d) Правило определения синонимов  
  
2. Что такое дерево зависимостей?  
**a) Графическое представление связей между словами в предложении**  
b) Способ классификации слов по частоте употребления  
c) Метод анализа морфологии слов  
d) Теория о семантике предложений  
  
3. В чем сущность гипотезы общего ядра?  
a) Идея о наличии общего языкового корня у разных языков  
b) Теория о синтаксической структуре языка  
c) Гипотеза о существовании универсальных языковых категорий  
**d) Предположение о происхождении языка**  
  
4. Дайте определение понятия семантическая просодия.  
**a) Исследование смысловых оттенков в языке**  
b) Анализ ритма и интонации в речи  
c) Изучение структуры предложений  
d) Оценка эмоциональной окраски текста  
  
5. Какие основные сферы применения корпусов?  
**a) Лингвистика, компьютерная лингвистика, психология**  
b) Медицина, искусственный интеллект, физика  
c) Химия, биология, математика  
d) История, археология, антропология  
  
6. Какие слова являются синонимами термина "аннотация"?  
**a) Резюме, выписка, содержание**  
b) Примечание, комментарий, объяснение  
c) Исследование, анализ, интерпретация  
d) Приложение, дополнение, документация  
  
7. Что такое конкорданс?  
a) Теория о структуре предложений  
b) Метод исследования морфологии слов  
**c) Список словоформ из текста с их контекстами**  
d) Графическое представление связей между словами  
  
8. Что такое корпус-конкордансер?  
**a) Программное обеспечение для работы с корпусами и конкордансами**  
b) Устройство для анализа семантики предложений  
c) Специалист по лингвистике корпусов  
d) Методика исследования дистрибуции слов  
  
9. Что такое дистрибуция?  
**a) Распределение слов в тексте по частоте употребления**  
b) Анализ морфологических характеристик слов  
c) Структура предложений в тексте  
d) Исследование синтаксических конструкций  
  
10. Что такое частность?  
**a) Частота употребления слова в тексте**b) Семантическая связь между словами  
c) Грамматическая категория слова  
d) Отношение слов к общему ядру  
  
11. Что такое морфологический анализатор?  
a) Методика анализа синтаксической структуры предложения)  
b) Специалист по изучению морфологии языка  
**c) Программа для автоматического определения грамматических характеристик слова**  
d) Теория о происхождении морфем в языке

Задание 2

Тест

Выберите один или несколько верных

вариантов ответа.

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?**

Впишите слово или число.

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 объем**.

10 Пропорциональное представление в корпусе текстов разных типов

называется «**балансирование**» корпуса.

Выберите единственный правильный

вариант ответа.

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

**1**.

Лингвистические особенности написания художественных произведений

 Особенности выражения эмоций в тексте романа «Отцы и дети» И. С. Тургенева:

1. Доминирование лексики, описывающей эмоциональные состояния героев.
2. Частотное использование описания эмоциональных кинем и просодем.
3. Использование языковой модели «называется выразитель эмоции (часть тела человека или орган) + лексическое обозначение испытываемой эмоции».
4. Богатый частеречный состав лексики, передающей эмоциональную информацию: существительные, прилагательные, глаголы, наречия, междометия.
5. Часто встречающаяся синтаксическая модель: «эмоциональное» наречие + глагол или глагол + «эмоциональное» наречие для трансляции эмоциональных переживаний героев.
6. Незначительное использование эмотивной лексики, преимущественно в диалогических высказываниях.

В романе «Отцы и дети» эмоциональная информация служит цели реализации эстетической функции всего произведения. Автор использует обширный лексический репертуар, чтобы создать необходимый эмоциональный фон, передать сложный мир чувств, настроений и стремлений героев, а также вызвать читательское сопереживание.

**2.**

baby’s bottle, car’s wheel, government’s decision

Что в британском(164), что в американском(172) языке эти словосочетания употребляются практически с одной частотой, т.к. они являются стандартными и распространенными, поэтому и в письменной, и в устной речи на этих двух языках употребляются одинаково часто.

**3**. Из анализа корпусов текстов видно, что обычно рефлексивный глагол с местоимением self употребляется в следующих форматах:  
1. Глагол + self: "He prides himself on his achievements."  
2. Местоимение + глагол + self: "My sister always blames herself for everything."  
3. 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.** Частота сочетаемости глагола help с инфинитивом в британском английском языке составляет примерно 2841. Это свидетельствует о том, что данная конструкция используется весьма часто.

**5.**

**Возьмем два глагола Beautiful и Handsome. Проанализируем корпуса текстов на английском языке. На основе проанализированного, мы можем сделать вывод о том, что Handsome чаще употребляется тогда, когда мы говорим о мужском поле или финансах(в значении щедрый), а beautiful употребляется тогда, когда мы говорим о женском поле или когда говорят о чем-то приятном(приятные манеры). Таким образом, мы видим, что синонимичные прилагательные в английском языке могут употребляться в разным ситуациях(частота употребления Beautiful будет больше, т.к. оно употребляется в большем количестве ситуаций.)**

**6.**

**9.**

Визуализация текста - это процесс представления текстовой информации в виде графических элементов, таких как диаграммы, графики, облака слов и т.д. Визуализация текста позволяет лучше понять структуру текста, выделить ключевые слова или фразы, а также обнаружить закономерности в распределении слов.  
  
Закон Ципфа - это явление, описывающее статистическую закономерность встречаемости слов в тексте. Согласно закону Ципфа, частота появления слов в тексте обратно пропорциональна их порядковому номеру по убыванию частоты. То есть самое часто встречающееся слово в тексте встречается вдвое чаще, чем второе по частоте, в три раза чаще, чем третье и т.д.  
  
Визуализация текста может помочь иллюстрировать закон Ципфа путем создания диаграммы частотности слов. На графике можно увидеть, как частота появления слов уменьшается с увеличением порядкового номера слова. Таким образом, визуализация текста может помочь подтвердить или опровергнуть закон Ципфа на конкретном корпусе текстов.

**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.

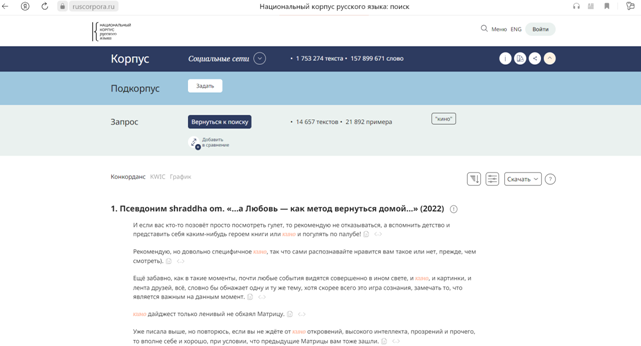
22. Контролируемые языки – способ пользователю системы машинного перевода и системе найти друг друга.

**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 статистических 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.**