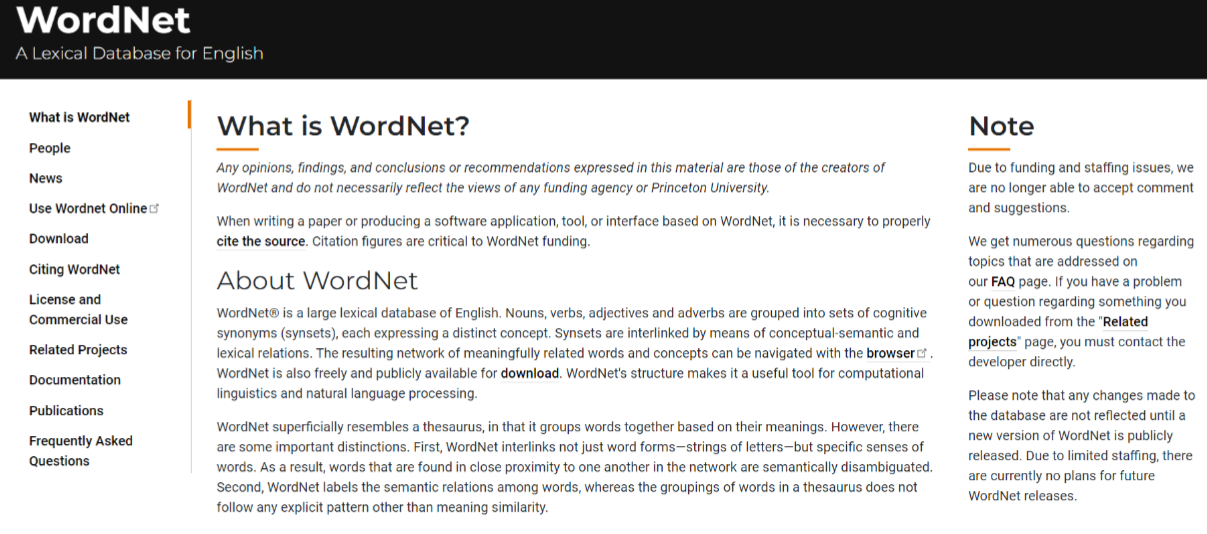
Уровень 1

1.

* WordNet (лексическая база данных)
* eLIBRARY.RU (электронная библиотека научных публикаций)
* Philology.ru (Справочные проф. ориентированные ресурсы)
* Google Books (поиск более 200,000,000,000 данных в американских и британских английских баз данных)
* Linguee (онлайн словарь)

WordNet

https://wordnet.princeton.edu/



1)Какая информация представлена?

WordNet представляет собой лексическую БД английского языка, которая организована в виде семантического сетевого словаря. Эта база данных содержит информацию о синонимах, антонимах, гипонимах и других семантических отношениях между словами, а также их определениях и примерах использования.

WordNet также предоставляет информацию о различных смыслах слов и их связях, что делает его полезным инструментом для компьютерных программ, работающих с обработкой естественного языка, а также для лингвистов и исследователей языка.

2) Для чего этот ресурс можно использовать?

1. В обработке естественного языка: WordNet помогает в разработке алгоритмов поиска информации, автоматического анализа текста, машинного перевода и других приложений в области обработки естественного языка.

2. В лингвистике: WordNet для изучения семантики слов, их синонимов, антонимов и других лексических отношений.

3. В обучении ИИ: WordNet может быть важным инструментом для разработки систем искусственного интеллекта, включая компьютерные программы, способные понимать и интерпретировать естественный язык.

4. В тезаурусах и словарях: WordNet может использоваться как основа для разработки тезаурусов, словарей и других лексических ресурсов.

3)Как производить запросы, синтаксис запроса

Основной способ использования WordNet - это через интерфейсы и инструменты, предоставляемые различными языками программирования и инструментами обработки естественного языка.

Например, если работать с Python, то можно использовать библиотеку NLTK (Natural Language Toolkit), которая предоставляет доступ к WordNet и позволяет выполнять запросы и операции над лексическими данными. Синтаксис запроса будет зависеть от синтаксиса и методов, предоставляемых выбранной библиотекой или интерфейсом.

4) Преимущества и недостатки

Преимущества ресурса WordNet:

1. Широкое покрытие: WordNet включает в себя обширную лексическую базу данных английского языка, охватывающую различные семантические отношения между словами.

2. Удобство использования: WordNet предоставляет доступ к разнообразной лексической информации, такой как синонимы, антонимы и т. д., что облегчает работу с лексическими данными.

3. WordNet широко используется в компьютерной лингвистике и обработке естественного языка для анализа текста, извлечения информации и других задач.

Недостатки ресурса WordNet:

1. Ограничение языка: WordNet в первую очередь ориентирован на английский язык, поэтому его покрытие других языков ограничено.

2. Отсутствие актуализации: Некоторые данные в WordNet могут быть устаревшими, поскольку он не всегда моментально отражает изменения в языке и лексических отношениях.

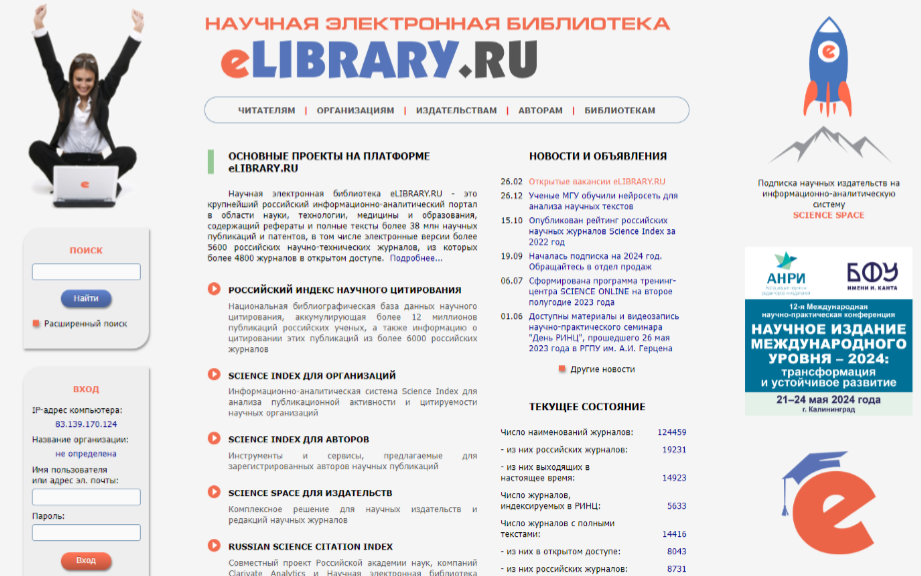
5)Ресурсы, похожие на WordNet:

1. Русский WordNet - аналогичный ресурс для русского языка, основанный на тех же принципах, что и англоязычный WordNet.

2. BabelNet - это ресурс, который объединяет семантические ресурсы на разных языках и предоставляет перекрестные ссылки между ними.

eLIBRARY.RU

https://elibrary.ru/defaultx.asp?



1)Какая информация представлена?

eLIBRARY.RU - это электронная библиотечная система, специализирующаяся на предоставлении доступа к научной литературе, журналам, книгам и исследованиям по различным областям науки и образования. представлена обширная коллекция научных публикаций и материалов, включая тематические журналы, книги, статьи, диссертации и другие научные исследования. Можно получить доступ к актуальным научным публикациям, провести поиск по различным тематикам, вести научные исследования и учиться новым знаниям.

Основные разделы включают в себя научные журналы, электронные книги, диссертации, статьи и конференции в различных областях знаний, таких как естественные науки, социальные науки, медицина

2) Для чего этот ресурс можно использовать?

- Научные исследования: можно получить доступ к актуальным научным публикациям, журналам, статьям и диссертациям

- Образование: Студенты, преподаватели и ученые могут использовать elibrary.ru для обогащения своих знаний, поиска материалов для учебы, подготовки курсовых работ и научных проектов.

- Профессиональное развитие: Специалисты различных отраслей могут использовать eLIBRARY.RU для поиска информации о последних тенденциях, новейших технологиях и отраслевых исследованиях.

- Публикации и авторство: Авторы могут публиковать свои исследования, статьи и материалы в eLIBRARY.RU, что способствует распространению научного знания и повышению видимости своих работ.

3)Как производить запросы, синтаксис запроса

Когда вы вводите поисковый запрос, убедитесь, что ключевые слова или фразы соответствуют тому, что хотите найти. Можно использовать различные операторы, такие как "AND" и "OR", для расширения или уточнения поискового запроса.

1. Простой запрос: "искусственный интеллект"

2. Расширенный запрос: "искусственный интеллект" AND "машинное обучение"

3. Включение альтернатив: "робототехника" OR "автономные системы"

4) Преимущества и недостатки

Преимущества ресурса eLIBRARY.RU:

1. Доступ к актуальным научным данным: eLIBRARY.RU предоставляет доступ к широкой коллекции научных публикаций, журналов, статей и книг по различным областям знаний.

2. Разнообразие тематик: eLIBRARY.RU охватывает различные области знаний, что делает его полезным для студентов, ученых, преподавателей, специалистов и всех, кто заинтересован в научной информации.

3. Возможность публикаций: Авторы могут использовать ресурс для публикации своих научных работ, что способствует распространению научного знания и обмену результатами исследований.

Недостатки ресурса eLIBRARY.RU:

1. Платный доступ: Некоторые материалы и функции доступны только для платных подписчиков, что может создавать ограничения для некоторых пользователей.

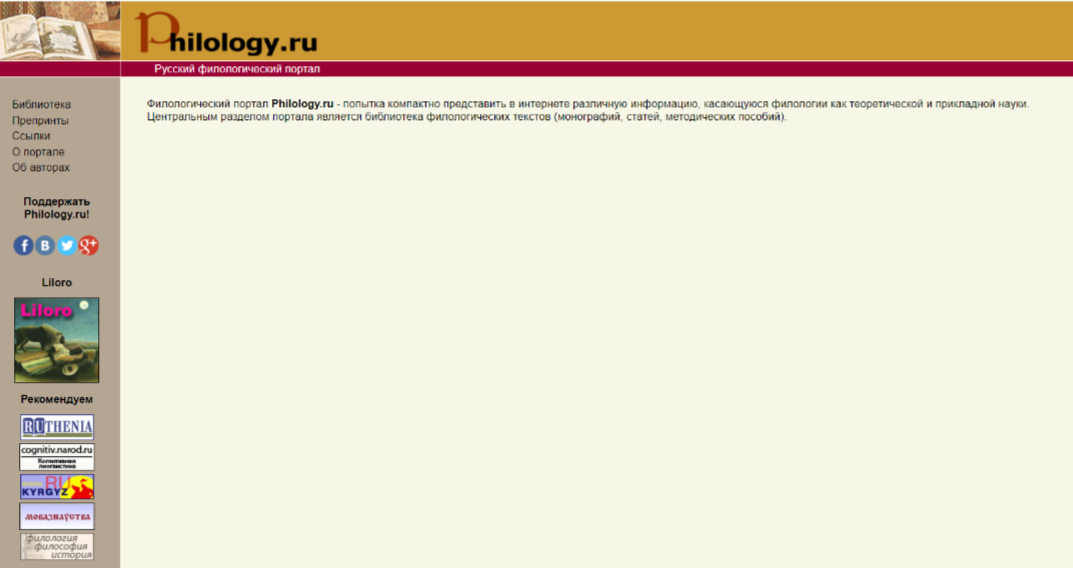
2. Ограниченный доступ к определенным изданиям: Не все издания и журналы могут быть доступны в eLIBRARY.RU из-за ограничений авторских прав или других причин.

5)Ресурсы, похожие на eLIBRARY.RU

1. CyberLeninka - это открытый цифровой архив, который предоставляет доступ к бесплатным полнотекстовым научным статьям в различных областях знаний.

2. ScienceDirect - это обширная электронная библиотека, которая обладает богатым содержанием научных журналов, статей, книг и баз данных в различных областях наук и технологий.

Philology.ru (http://www.philology.ru/)



1)Philology.ru - это онлайн-платформа, предоставляющая информацию о лингвистике, филологии, иностранных языках, литературе и культуре. Ресурс содержит информацию о научных статьях, книгах, методических пособиях, словарях, толкованиях, культурных материалах и многом другом, связанном с указанными областями.

2)Этот ресурс можно использовать для:

- Поиска информации и литературы по филологии и лингвистике;

- Углубленного изучения вопросов языка, литературы и культуры;

- Поиска научных статей и исследований;

- Подготовки к учебе, написанию научных и учебных работ.

3)Для выполнения запросов на ресурсе можно использовать поиск по ключевым словам, фразам, авторам, названиям книг и статей. Пример синтаксиса запроса: "язык", "лингвистика", "филология", "литература", "русский язык" и т.д.

4)Преимущества:

- Широкий спектр информации по филологической области;

- Возможность доступа к научным статьям и методическим пособиям;

- Проверенный и авторитетный источник информации.

Недостатки:

- Не всегда содержит самую актуальную информацию;

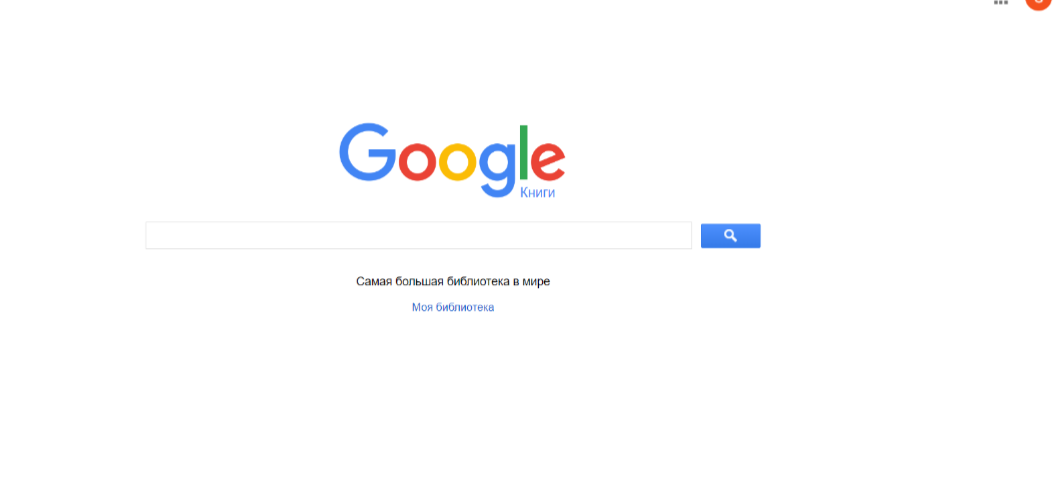
- Не всегда имеет информацию на всех языках.

5)Похожие ресурсы

1)[gramota.ru](https://www.similarweb.com/ru/website/gramota.ru/)

2)[iling-ran.ru](https://www.similarweb.com/ru/website/iling-ran.ru/)

Google Books(https://books.google.com/?hl=ru)



1)Google Books представляет собой онлайн-библиотеку, содержащую множество электронных книг различных жанров, тематик и языков. Ресурс предоставляет доступ к отрывкам и целиком книг, а также каталогизирует миллионы публикаций.

2)Этот ресурс можно использовать для:

- Поиска и чтения электронных книг;

- Поиска цитат, отрывков и информации из книг;

- Исследовательских работ;

- Изучения различных тем и областей.

3)Для производства запросов можно использовать строку поиска на главной странице Google Books. Синтаксис запроса аналогичен поиску в общем поисковике Google.

4)Преимущества:

- Большой объем электронных книг различных жанров и тематик;

- Возможность просмотра отрывков книг перед покупкой или чтением;

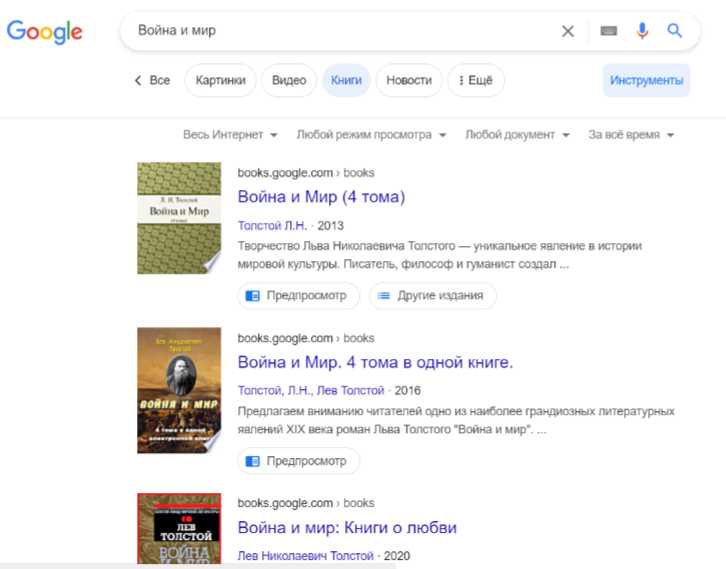
- Каталогизация и поиск информации внутри книг.

Недостатки:

- Не всегда доступны полные версии книг из-за авторских прав;

- Не все книги доступны для определенных регионов;

- Не всегда наличие актуальной и современной литературы.



Linguee (https://www.linguee.ru/)



1)Linguee - это онлайн-словарь и мультиязычный поисковик, который предоставляет переводы и примеры использования слов и фраз на различных языках. Ресурс также содержит большую базу фразеологизмов, толкований и примеров перевода с учетом контекста.

2)Этот ресурс можно использовать для:

- Поиска переводов слов и фраз с одного языка на другой;

- Поиска контекстуальных примеров использования слов;

- Поиска специализированных терминов и фразеологизмов.

3)Для производства запросов можно воспользоваться строкой поиска на главной странице Linguee. Синтаксис запроса обычно включает слово или фразу на исходном языке, за которым следует переводимое слово или фраза на целевом языке.

4)Преимущества:

- Большая база слов и фраз на различных языках;

- Предоставление контекстуальных примеров использования слов и фраз;

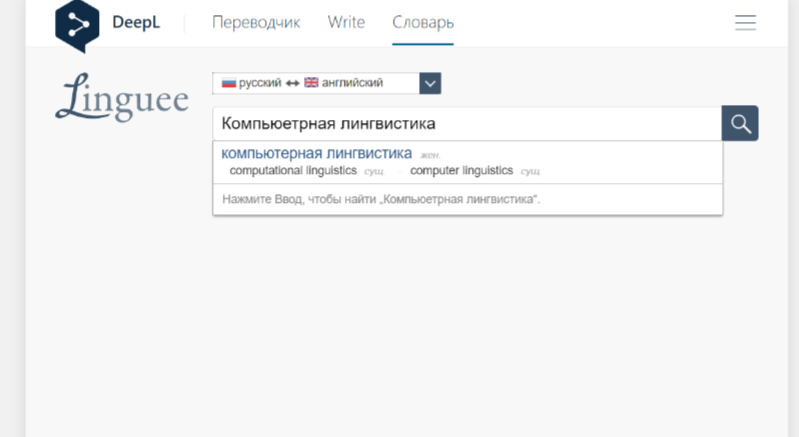
- Возможность поиска специализированных терминов и выражений.

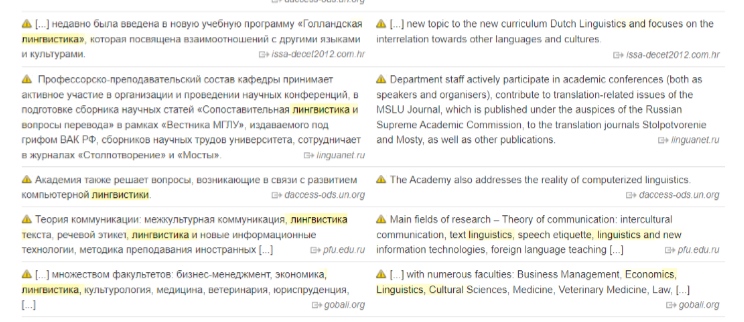
Недостатки:

- Не всегда содержит полный спектр возможных переводов;

- В некоторых случаях может отсутствовать более адекватный перевод для определенных слов и выражений;

- Не всегда могут быть представлены последние изменения в языке.





5)Похожие ресурсы

1. Reverso Context: Этот ресурс предоставляет переводы и контекстуальные примеры на нескольких языках, помогая пользователям понять, как использовать слова и фразы в правильном контексте.

2. ABBYY Lingvo: Этот ресурс предлагает онлайн-словарь и переводчик, а также функцию поиска по тексту для быстрого доступа к переводам и толкованиям слов.

2.

Вопросы для обсуждения:

1. Единицей корпуса являются тексты или их достаточно значительные фрагменты, включающие, например, какие-то полные фрагменты макроструктуры текстов данной проблемной области.
2. Корпусы составляют в зависимости от поставленных целей и задач, они должны отражать текстовое разнообразие(если это необходимо), иметь специализированность или общую направленность текстов.
3. Исследовательскими называются такие корпусы, которые предназначены преимущественно для изучения различных аспектов функционирования языковой системы. Они строятся не post factum — после проведения какого-либо исследования, а до его проведения.   
   Статический корпус – корпус, отражающий определенное временное состояние языковой системы.   
   Параллельный корпус — это двуязычный корпус, то есть текст оригинала и его перевод на какой-то другой язык, причем эти два текста не просто лежат рядом друг с другом, а должны быть выровнены: отдельные фрагменты оригинала должны совпадать с соответствующими фрагментами перевода.

|  |  |
| --- | --- |
| **Критерии классификации** | **Виды корпусов** |
| Тип языковых данных | Устные / Письменные / Смешанные |
| Формат представления данных | Мономодальные / Мультимодальные (мультимедийные) |
| Объем и полнота данных | Общие (представительные, национальные) / Для специальных целей: мониторинговые, иллюстративные |
| Число представленных языков | Одноязычные / Двуязычные / Многоязычные |
| Число языков в одном тексте (для дву- и многоязычных корпусов) | Параллельные / Смешанные |
| Выравнивание текстов на разных языках (для дву- и многоязычных корпусов) | Соизмеримые или параллельные / Несоизмеримые или непараллельные |
| Наличие единых принципов отбора языкового материала | Сопоставимые / Несопоставимые |
| Представленность языкового материала | Полнотекстовый / Фрагментированный (в т.ч. корпуса n-грамм и конкордансы) |
| Временной период, к которому относятся собранные языковые данные | Синхронные / Диахронные |
| Наличие разметки | Размеченные (аннотированные) / Неразмеченные (неаннотированные) |
| Тип разметки (для размеченных корпусов) | Метатекстовая разметка / Лингвистическая разметка (морфологическая, синтаксическая, семантическая) / Экстралингвистическая разметка |
| Доступность | Свободно доступные / Частично доступные / Закрытые (коммерческие) |

1. Британский национальный корпус: содержит около 100 миллионов словоупотреблений, собранных из различных текстов на английском языке.  
   По виду корпусов:   
   По типу текстов: В его состав входят тексты разных жанров и стилей.  
   По разнообразию языков: Корпус включает в себя множество форм английского языка, что можно использовать для исследований(Многоязычные)  
   По временному периоду: представлены тексты разных периодов(Диахронные)
2. Глоссарий по теме «Корпусная лингвистика» :  
   ***Конкорданс*** - алфавитный перечень всех слов какого-либо текста с указанием контекстов их употребления; тип словаря, представленный в таком виде  
   ***Рандомизация*** - процедура случайного выбора элементов статистической совокупности при проведении выборочного исследования  
   ***Коллокация*** - в корпусной лингвистике под коллокацией понимается последовательность слов или терминов, частотность совместного появления которых в корпусе выше, чем ожидаемая вероятность их совместного появления.  
   ***Подмассив*** - это любая последовательность чисел в массиве (непрырывная).  
   ***Парсинг*** - это синтаксический анализ, который автоматически производится парсером – специальной программой или скриптом.  
   ***Лемматизация*** - процесс привода словоформы к лемме — её нормальной (словарной) форме.  
   ***Корпус- менеджер* -** программное обеспечение для корпуса**,** котороеподсчитывает конкретные словоформы, группы словосочетаний и также может выводить результаты статистики, представить в виде диаграммы.
3. Среди обнаруженных мною ресурсов интерес представляют 2:  
   первый - <http://blogs.it-claim.ru/izelentsov/>, где представленны 4 корпуса, созданные группами специалистов.   
   второй - <https://elementy.ru/catalog?type=127&ysclid=ltn8egrg4e692471229>, где собрана информация о корпусах

**Лабораторная работа 10**

1. Результаты представленные в таблице могут свидетельствовать о том, что наименьшее количество вхождений ключевого слова в Британском корпусе

|  |  |  |  |
| --- | --- | --- | --- |
|  | НКРЯ | COCA | BNC |
| Количество вхождений | 39 076(315 677) | 49303 | 5370 |

|  |  |  |  |
| --- | --- | --- | --- |
| № Примера | НКРЯ | COCA | BNC |
| 1 | Я не люблю *русский* рок, но эта одна из тех 5-6 групп, которые я слушаю [Переписка в icq между agd-ardin и Колючий друг (08.02.2008)] | It's just very expensive. # As for the mirror experiment, the **Russian** version failed rather quickly because they used a robotic lander, and it was unableх[Moon hoax: why not use telescopes to look at the landers? | Bad] | Here is a description by Diderot of a [**Russian**](http://bncweb.lancs.ac.uk/cgi-binbncXML/context.pl?thin=0&numOfFiles=859&inst=50&qname=dercomputer_1710219560&max=108&numOfSolutions=5370&queryID=dercomputer_1710219560&thMode=M5370%23859%23no_subcorpus%23%23&theID=dercomputer_1710219560&theData=%5Bword%3D%22Russian%22%25c%5D&chunk=1&queryMode=simple&view2=nonrandom&simpleQuery=Russian&program=search&view=list&listFiles=0&phon=0&subcorpus=no_subcorpus&queryType=CQL&qtype=0&text=A04&refnum=2&theShowData=Russian&len=-24&showTheTag=0&color=0&begin=1358&token_offset=8&nodeCount=1&hitSunit=1358&spids=1&interval=11&urlTest=yes) country scene by Le Prince, from criticism of the Salon in 1767. [Art criticism: a user's guide.] |
| 2 | Мое сознание постепеннно забыло что *русский* — это мой язык " by default " [Запись LiveJournal (2004)] | Now what we we talking about? # Although you can say that the **Russian** mirror performance is poorer than the US ones in some respects, as described above[Moon hoax: why not use telescopes to look at the landers? | Bad] | The shrugger doesn't care whether he lives or dies — and the designer duel arranged for himself by this divided and indifferent man is a form of [**Russian**](http://bncweb.lancs.ac.uk/cgi-binbncXML/context.pl?thin=0&numOfFiles=859&inst=50&qname=dercomputer_1710219560&max=108&numOfSolutions=5370&queryID=dercomputer_1710219560&thMode=M5370%23859%23no_subcorpus%23%23&theID=dercomputer_1710219560&theData=%5Bword%3D%22Russian%22%25c%5D&chunk=1&queryMode=simple&view2=nonrandom&simpleQuery=Russian&program=search&view=list&listFiles=0&phon=0&subcorpus=no_subcorpus&queryType=CQL&qtype=0&text=A05&refnum=3&theShowData=Russian&len=-32&showTheTag=0&color=0&begin=933&token_offset=28&nodeCount=1&hitSunit=933&spids=1&interval=11&urlTest=yes) roulette. [Authors] |
| 3 | От самого легкого — к самому трудному: английский, немецкий, *русский*, французский (с транслита) [Наши дети: Подростки (2004)] | mith &; Nephew countered with its own experts, including a top deputy of the **Russian** Federation's Department of Libraries, flown into Houston to be deposed. # KCI[The bloody patent battle over a healing machine - Fortune] | Even so, he wrote, what might, à la rigueur , be acceptable for a [**Russian**](http://bncweb.lancs.ac.uk/cgi-binbncXML/context.pl?thin=0&numOfFiles=859&inst=50&qname=dercomputer_1710219560&max=108&numOfSolutions=5370&queryID=dercomputer_1710219560&thMode=M5370%23859%23no_subcorpus%23%23&theID=dercomputer_1710219560&theData=%5Bword%3D%22Russian%22%25c%5D&chunk=1&queryMode=simple&view2=nonrandom&simpleQuery=Russian&program=search&view=list&listFiles=0&phon=0&subcorpus=no_subcorpus&queryType=CQL&qtype=0&text=A08&refnum=6&theShowData=Russian&len=-56&showTheTag=0&color=0&begin=512&token_offset=17&nodeCount=1&hitSunit=512&spids=1&interval=11&urlTest=yes) Grandmaster seems merely perverse in a retired artist. [The big glass] |

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

|  |  |  |  |
| --- | --- | --- | --- |
| № примера | НКРЯ | COCA | BNC |
| 1 | из россии | российская версия | принадлежность к россии |
| 2 | язык | российская версия | устойчивое выражение |
| 3 | язык | название страны | принадлежность к стране |

1. В моем случае все взятые слова с точки зрения морфологии являются - прилагательными, а с лексической точки зрения они различаются: где-то это отношение к стране, где-то выражение и т.д.
2. Рассмотренные корпуса можно использовать для сравнения частоты появления слова в текстах данных корпусов, сравнение их лексических значений и многое другое.

Уровень 2

1.

|  |  |
| --- | --- |
| **“Корпусная лингвистика”** | |
| В чем сущность эмпирической методологии  в лингвистическом анализе? | Эмпирическая методология в лингвистическом анализе основывается на исследовании языка на основе фактических данных, полученных путем наблюдения, экспериментов и анализа языковых явлений в реальных ситуациях. Сущность этого подхода заключается в том, что он ставит целью собирать и анализировать конкретные эмпирические данные для выявления общих закономерностей и особенностей языковых явлений. |
| Дайте определение понятию коллокация (collocation) | в корпусной лингвистике под коллокацией понимается последовательность слов или терминов, частотность совместного появления которых в корпусе выше, чем ожидаемая вероятность их совместного появления. |
| Сформулируйте закон Ципфа | Закономерность распределения частоты слов естественного языка. Если все слова языка или длинного текста упорядочить по убыванию частоты использования, частота n-го слова в списке окажется обратно пропорциональной его порядковому номеру n.(Самое распространенное слово в тексте встречается: в два раза чаще, чем второе по частоте слово; в три раза чаще, чем следующее слово; и далее до наименее частого.) |
| Дайте определение понятию дискурсивной просодии (discourse prosody) | Дискурсивная просодия — это область исследований, занимающаяся изучением интонационных, темпоральных и других акустических особенностей речи, которые несут смысловую нагрузку в рамках дискурса. Данное понятие включает в себя анализ изменений в интонации, ритме, громкости, паузах и других акцентуальных элементах, которые присущи устной речи и способствуют передаче информации и эмоциональной окраске высказывания. Дискурсивная просодия помогает понять, каким образом тон, интонация и другие звуковые аспекты речи влияют на восприятие и интерпретацию высказывания в контексте дискурса. |
| Дайте определение понятию семантической просодии (semantic prosody) | Семантическая просодия (semantic prosody) - это концепция в лингвистике, которая относится к тому, как определенные слова или конструкции носят определенный оттенок эмоций, оценки или ассоциаций в контексте их использования. |
| В чем сущность гипотезы общего ядра (common core hypothesis)? | Согласно гипотезе общего ядра, существует набор языковых единиц и структур, которые присутствуют во всех языках и являются базовыми для языкового выражения мысли. Эти общие элементы часто связаны с универсальными когнитивными механизмами и способами представления информации, что делает их более стабильными и широко применимыми в языковых системах. |
| Что такое дерево зависимостей? | Деревья зависимостей — наиболее наглядный и наиболее распростра­ненный способ представления синтаксической структуры предложения. При этом предложение представляется как линейно упорядоченное множество элементов (словоформ), на котором можно задать ориентированное дерево (узлы — элементы множества). |
| Постройте дерево зависимостей для простого предложения английского языка | Предложение: "He reads a book."  - he (субъект) -> reads (глагол)  - a (определение) -> book (существительное)  - reads (глагол) -> he (субъект)  - reads (глагол) -> book (дополнение) |
| Постройте дерево зависимостей для сложного предложения английского языка | Предложение: "She likes the movie that I watched yesterday."  - she (субъект) -> likes (глагол)  - the (определение) -> movie (существительное)  - likes (глагол) -> she (субъект)  - likes (глагол) -> movie (дополнение)  - watched (причастие) -> I (подлежащее)  - watched (причастие) -> yesterday (обстоятельство времени)  - movie (существительное) -> that (относительное местоимение)  - that (относительное местоимение) -> watched (причастие) |
| Назовите основные сферы применения корпусов? Дайте синонимы термину аннотация (annotation) | Корпуса используются в основном для лингвистического анализа, обучения и тестирования моделей машинного обучения, изучения языков и прочих исследований. Синонимы термина «аннотация» включают «описание», «разметка» и «обозначение». |
| Дайте определение понятию репрезентативность корпуса. | Репрезентативность корпуса - это степень, в которой корпус отражает целевую популяцию или явление. Чем более репрезентативен корпус, тем точнее он отражает разнообразие и структуру исследуемого языка. |
| Дайте определение понятию кластерный анализ (cluster analysis) | Кластерный анализ (Cluster Analysis) - это метод статистического анализа, который используется для выявления групп (или “кластеров”) схожих элементов в наборе данных. Он позволяет сегментировать данные на основе их сходства, что позволяет выявлять общие паттерны или взаимосвязи между элементами. |
| Что такое конкорданс? | алфавитный перечень всех слов какого-либо текста с указанием контекстов их употребления; тип словаря, представленный в таком виде |
| Что такое конкордансер? | это инструмент или программа, которая позволяет находить все вхождения определенного слова или выражения в тексте и предоставлять информацию о контексте, в котором они встречаются |
| Дайте определение понятию дистрибуция | Дистрибуция в лингвистике (позиционное распределение) — это множество всех окружений (контекстов), в которых встречается некоторый элемент, то есть множество всех (различных) возможных позиций элемента относительно позиций других элементов. |
| Покажите дистрибуцию двух синонимов (на выбор) | счастливый и радостный синонимы, но при этом у них разные множества контекстов |
| Дайте определение понятию частотность | Частотность в лингвистике обозначает количественную меру того, насколько часто какой-либо лингвистический элемент (слово, звук, грамматическая конструкция и т. д.) встречается в данном корпусе текста или языковом материале |
| Определите частотность любой лексической единицы английского языка по различным корпусам | Например в Британском национальном корпусе слово “help” встречается 36884 раз |
| Дайте определение понятиям лемма, лемматизация | ***Лемматизация*** - процесс привода словоформы к лемме — её нормальной (словарной) форме. Лемма - начальная форма слова |
| Дайте определение понятию лексическая плотность | это показатель, характеризующий степень насыщенности текста лексическими единицами. Она отражает количество лексических единиц в тексте или в его отдельной части, сравнительно с общим числом слов |
| Что такое морфологический анализатор? | Морфологический анализатор – это инструмент, определяющий грамматические характеристики слова |
| Дайте определению понятию сегментация корпуса | Разделение корпуса на сегменты |

2.

"[Что такое корпус. Тест](https://edu.vsu.ru/mod/resource/view.php?id=919012)"

1. Тексты, составляющие лингвистический корпус,

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,

она должна отражать состояние языка в определенный период времени.

она должны быть размечена.

4. Обязательными свойствами хорошего лингвистического корпуса являются

представительность.

присутствие разметки.

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

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 тип корпуса, по которому велся поиск.

7. Согласно материалам лекции, основная часть Национального корпуса русского языка включает 374 449 975 слов.

8. Количество найденных результатов обозначается в НКРЯ как количество «словоформ».

9. Нормализация значений обычно предполагает вычисление отношения количества вхождений к объему корпуса, умноженного на коэффициент, корректирующий частоту встречаемости слова в корпусе.

10. Пропорциональное представление в корпусе текстов разных типов называется «репрезентативность» корпуса.

**Выберите единственный правильный вариант ответа.**

11. Что означает понятие «накопительная сбалансированность»?

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

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

Она проведена в соответствии со школьной грамматикой.

13. Чем отличается лингвистический корпус от электронного словаря?

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?

Google Books Ngram Viewer

"[Разметка корпусов. Тест](https://edu.vsu.ru/mod/resource/view.php?id=919020)"

**Выберите один или несколько верных вариантов ответа.**

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.

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.

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

4. Автоматическая разметка

начинается с деления текста на минимальные фрагменты.

предполагает отнесение словоформы к начальной форме.

позволяет быстро провести аннотирование по частям речи.

5. Экстралингвистическая разметка

предполагает демонстрацию сведений об авторе.

в НКРЯ выражается в виде «паспорта текста».

6. Морфологическая разметка в НКРЯ

произведена с помощью компьютерных программ.

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.

8. Семантическая разметка в НКРЯ

помечает семантические признаки грамматических разрядов.

помечает семантические признаки словообразовательных показателей.

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.

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

двум леммам.

**Впишите слово.**

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 «удаление шума».

13.Парсер – это компьютерная программа, производящая синтаксическую разметку.

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

15. Источником автоматической семантической разметки часто служат компьютерные программы.