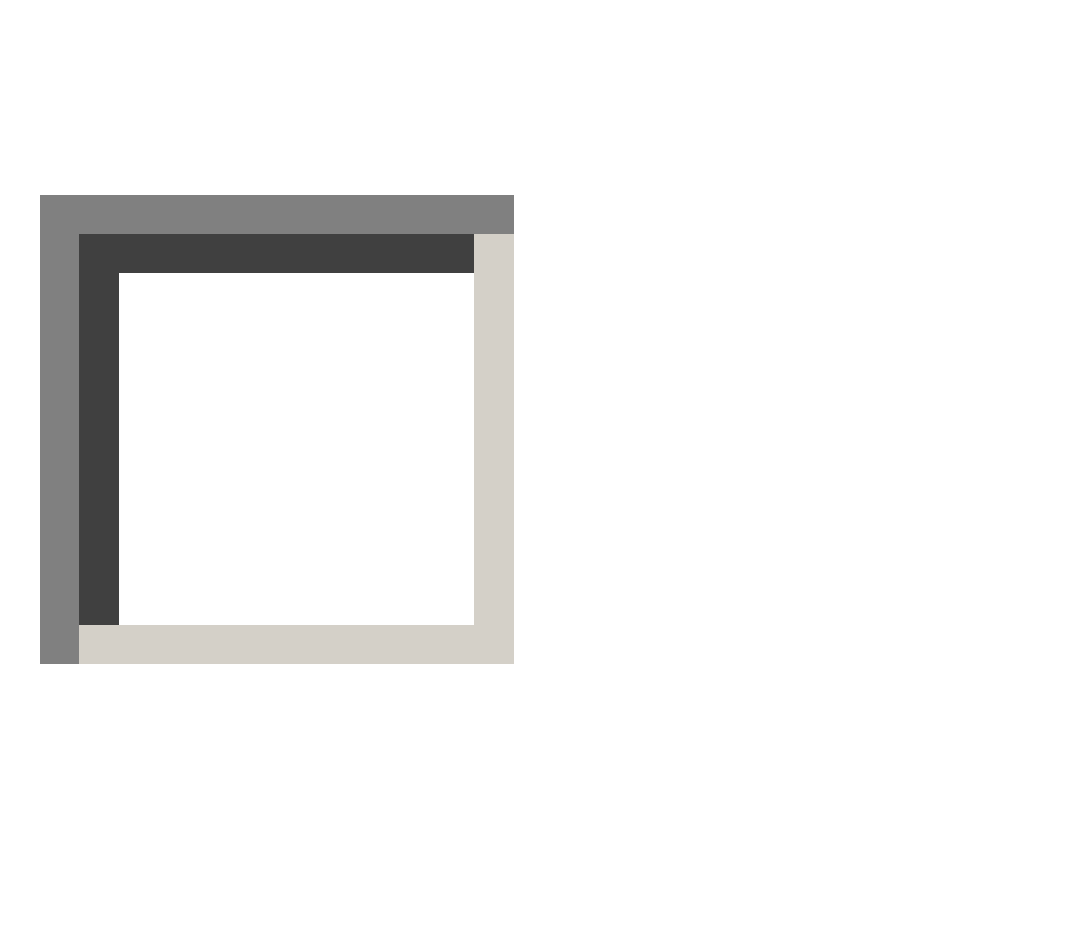
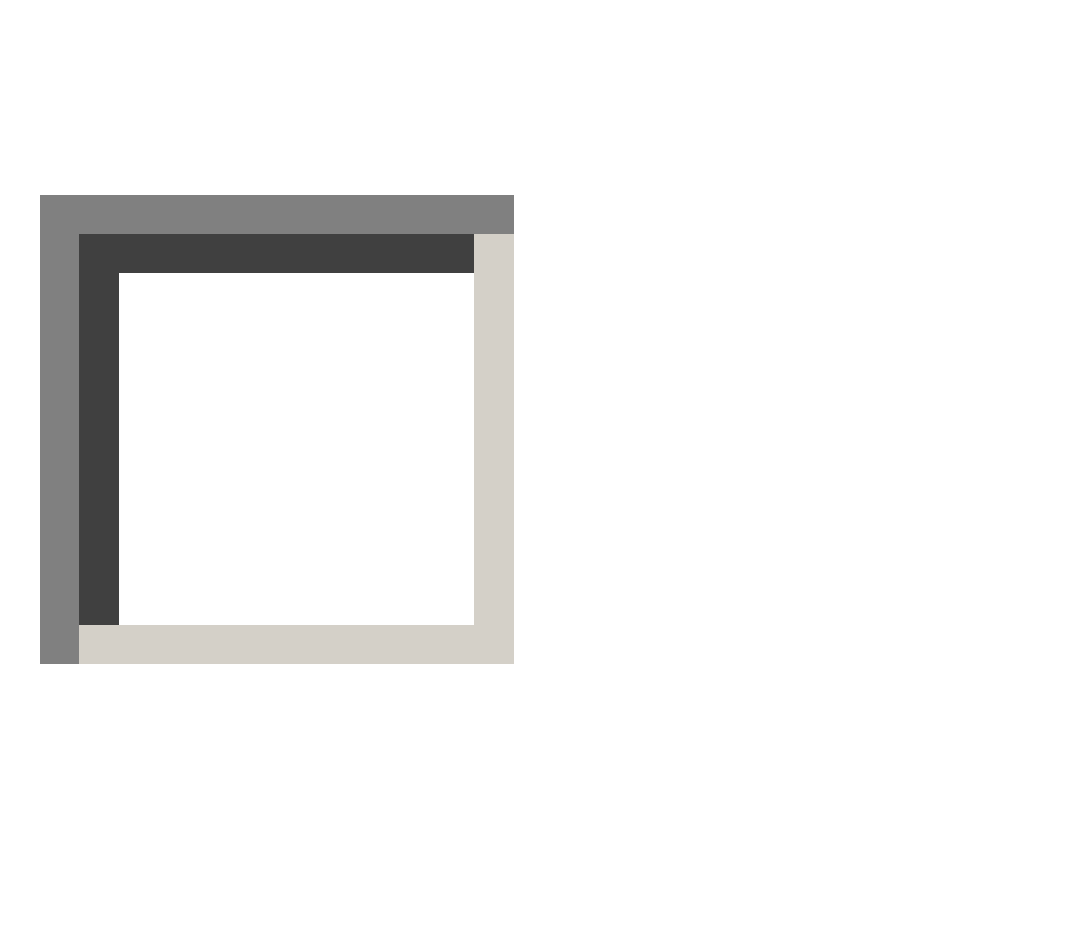
| **“Корпусная лингвистика”** | |
| --- | --- |
| В чем сущность эмпирической методологии  в лингвистическом анализе? | Эмпирическая методология в лингвистическом анализе основывается на исследовании языка на основе фактических данных, полученных путем наблюдения, экспериментов и анализа языковых явлений 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 |

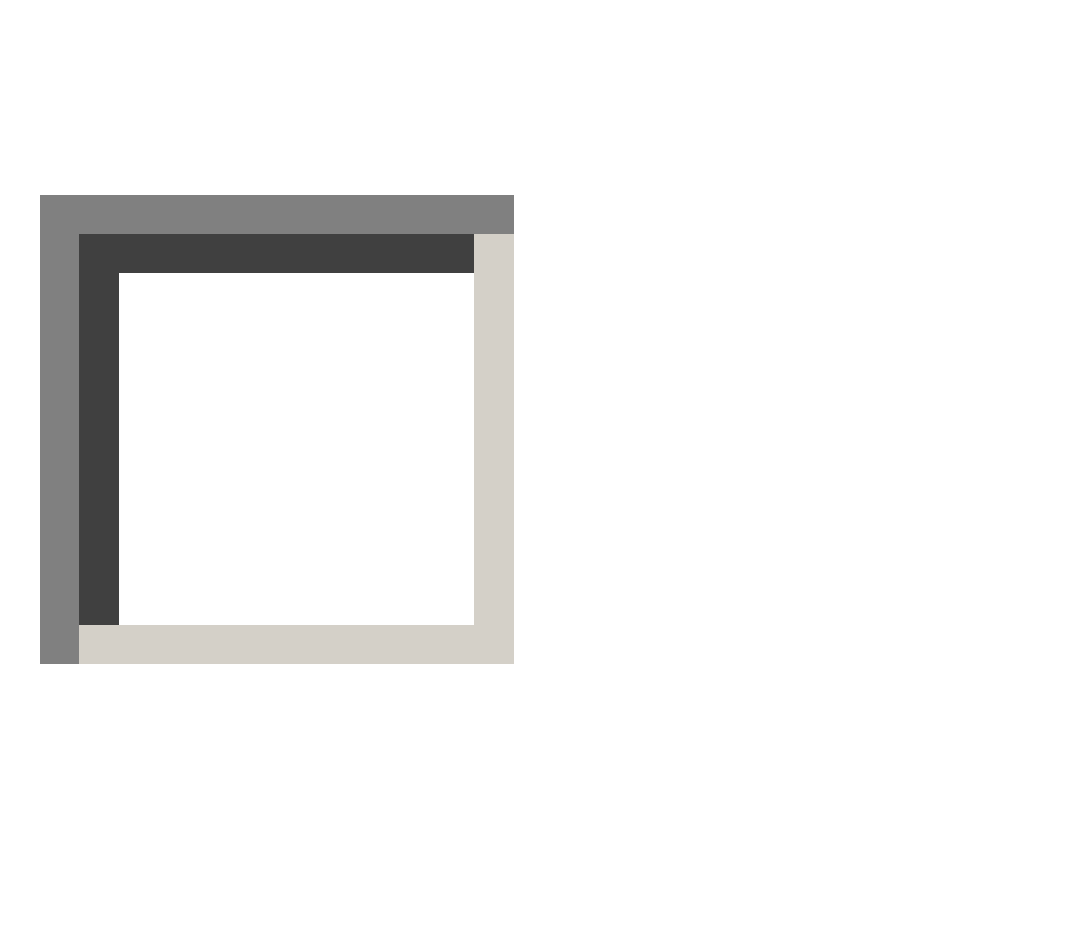
Тест “Что такое корпус”**.**

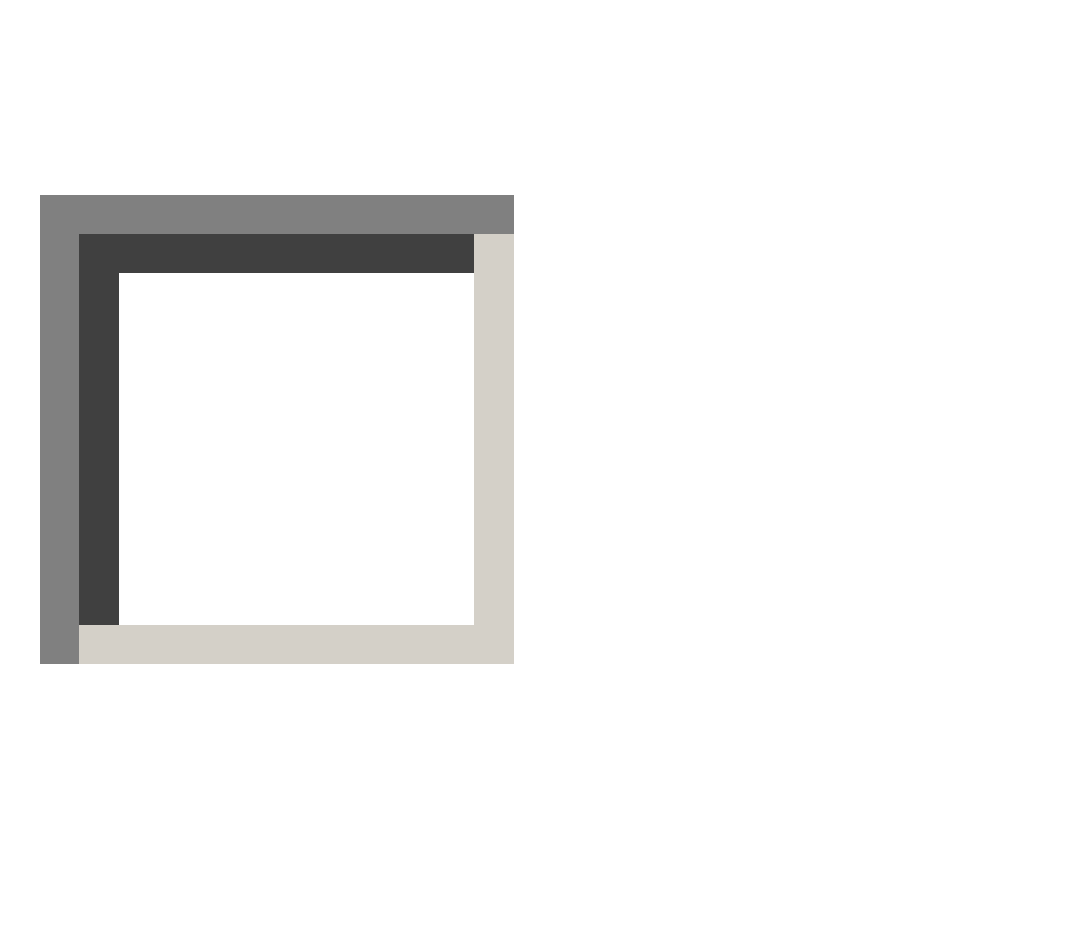
1. Тексты, составляющие лингвистический корпус,

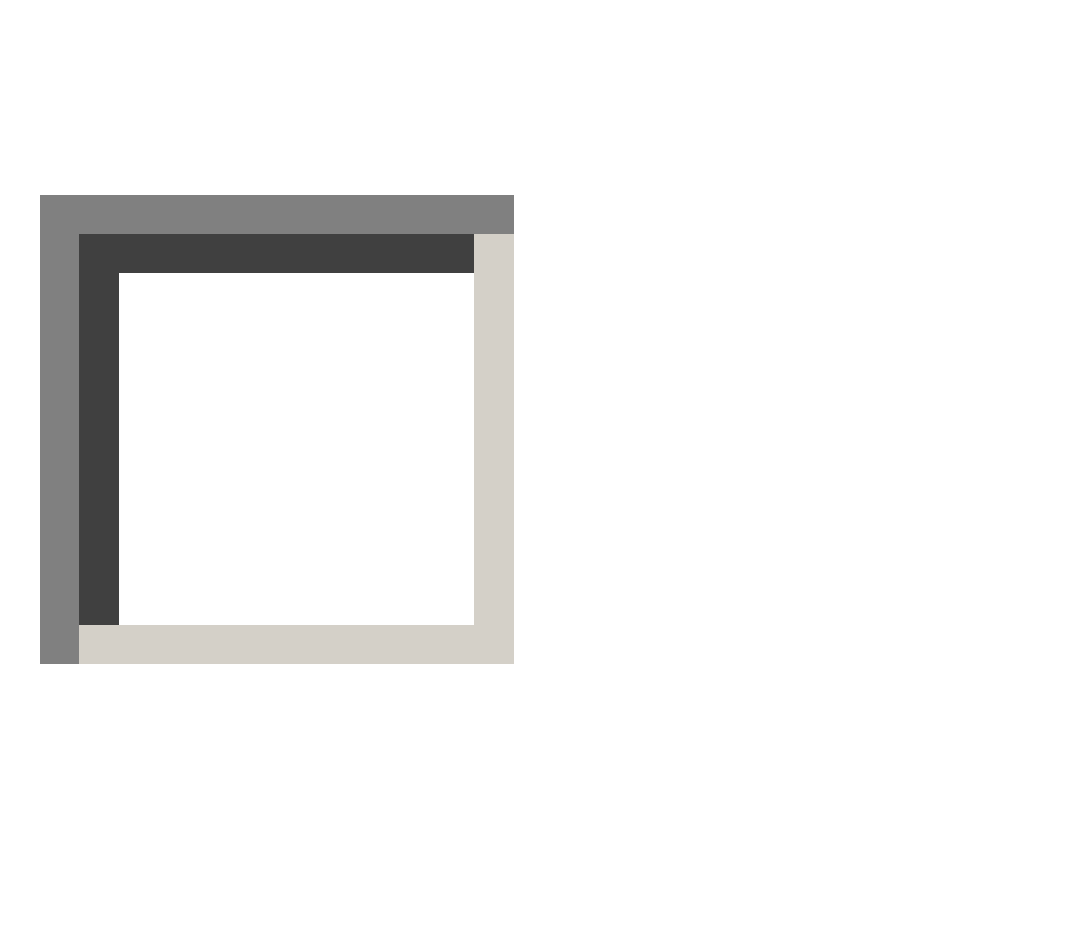
размечены лингвистами и программистами.

2. В лингвистический корпус могут входить

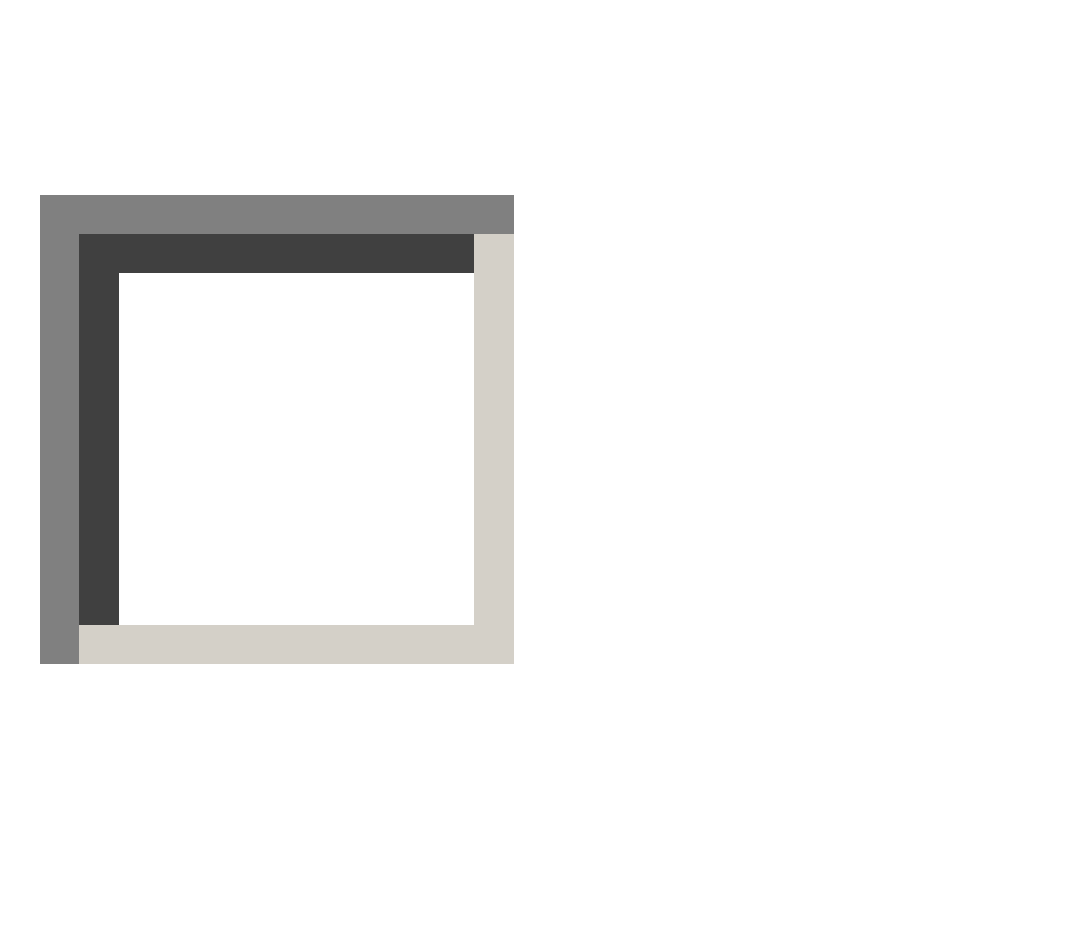
тексты электронной коммуникации.

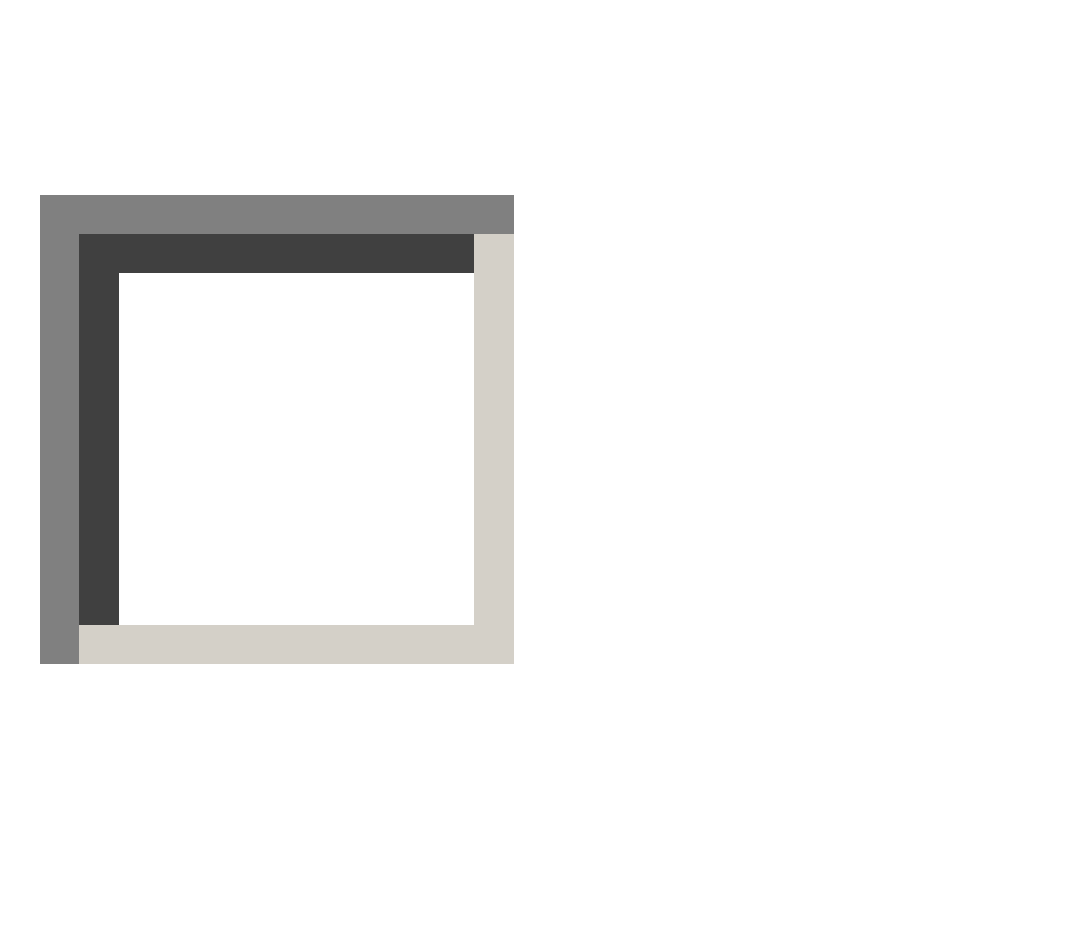
записи устной речи детей.

транскрипции теле- и радиопередач.

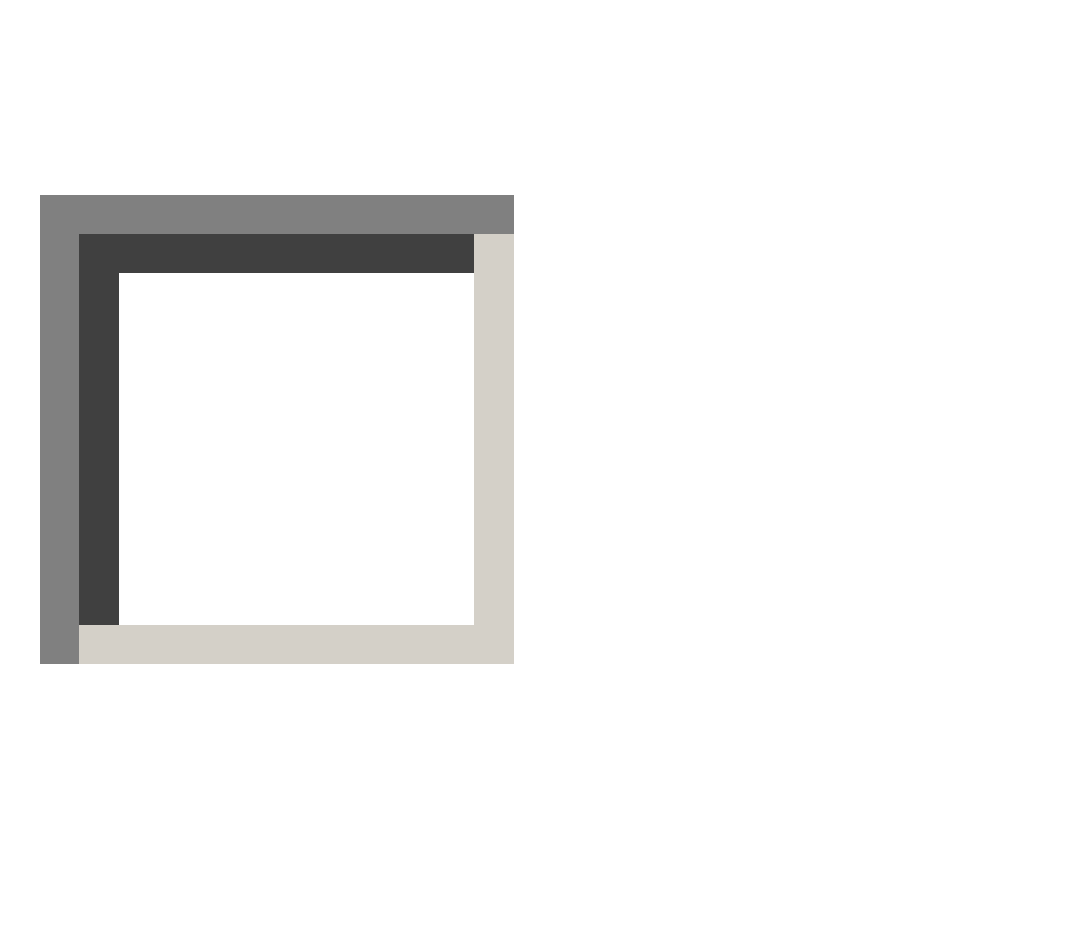
произведения художественной литературы.

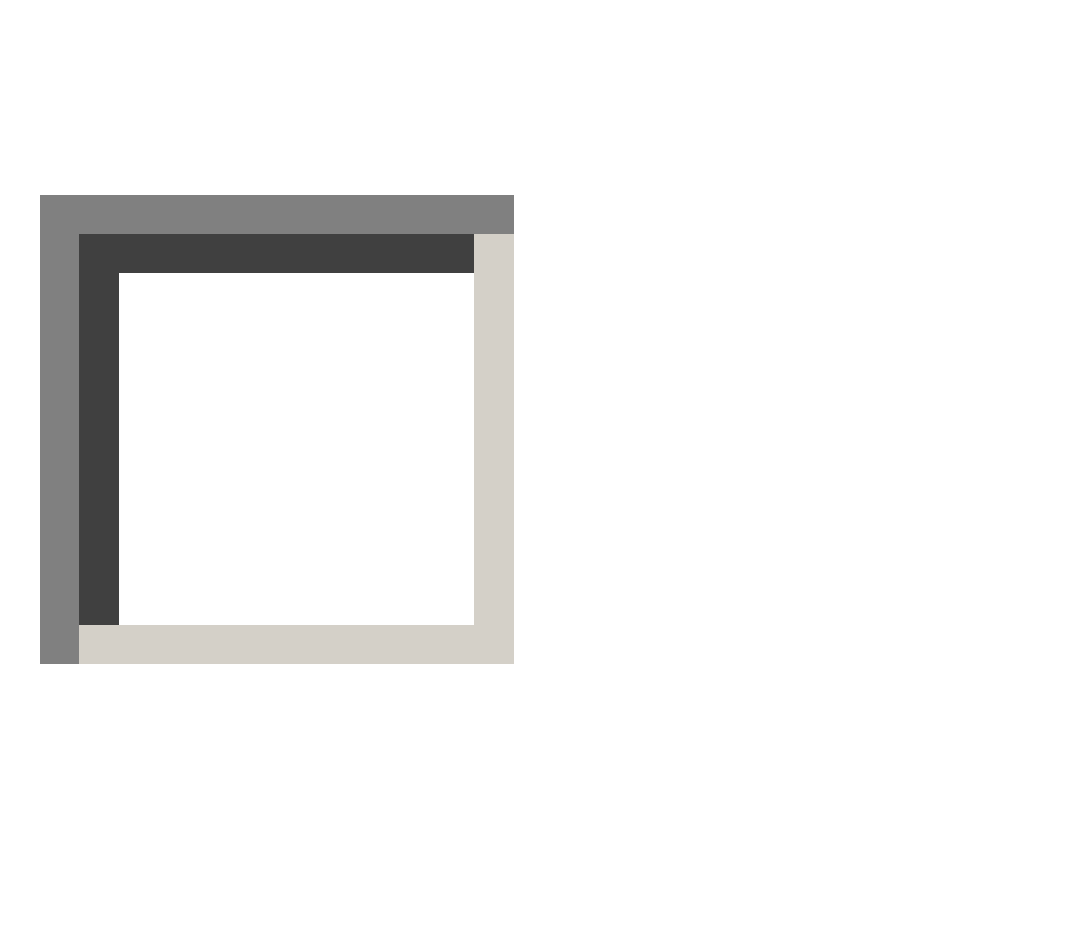
3. Для того, чтобы электронная коллекция текстов могла считаться корпусом,

она должна отражать состояние языка в определенный период времени.

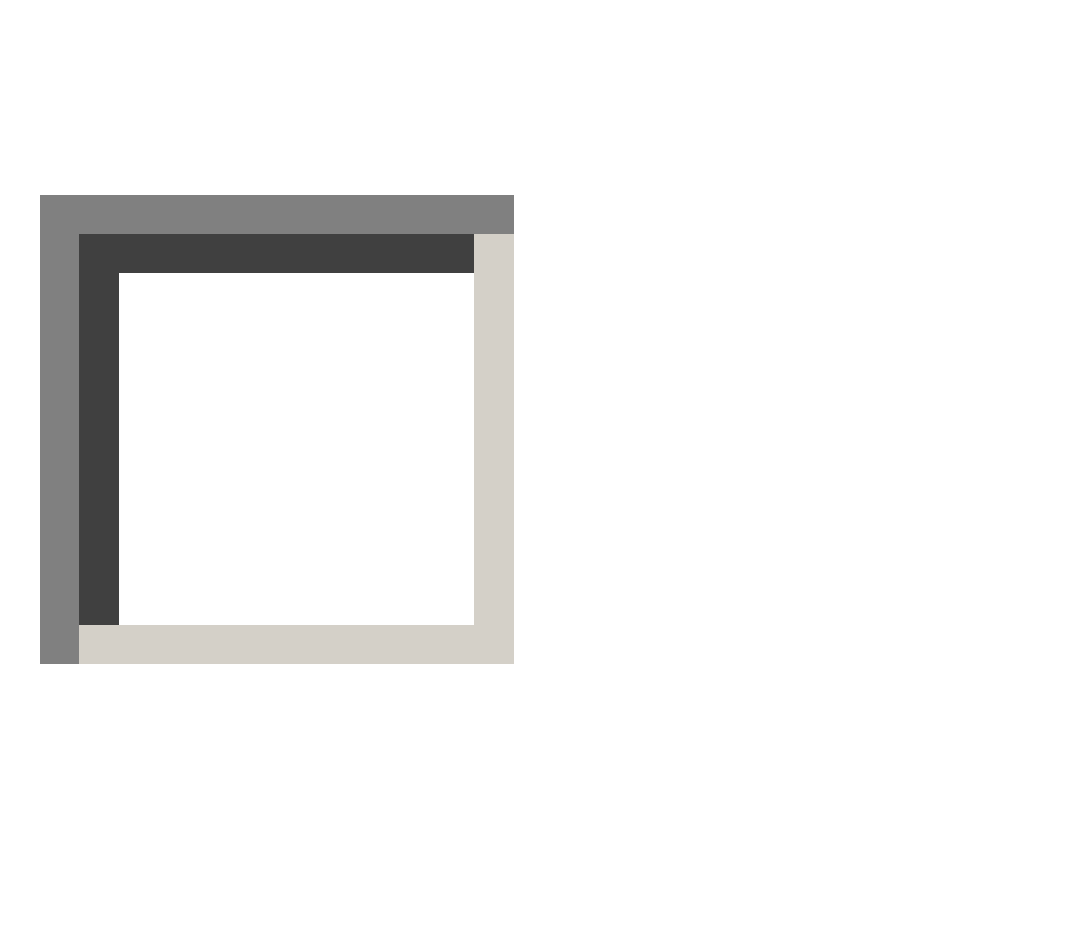
она должны быть размечена.

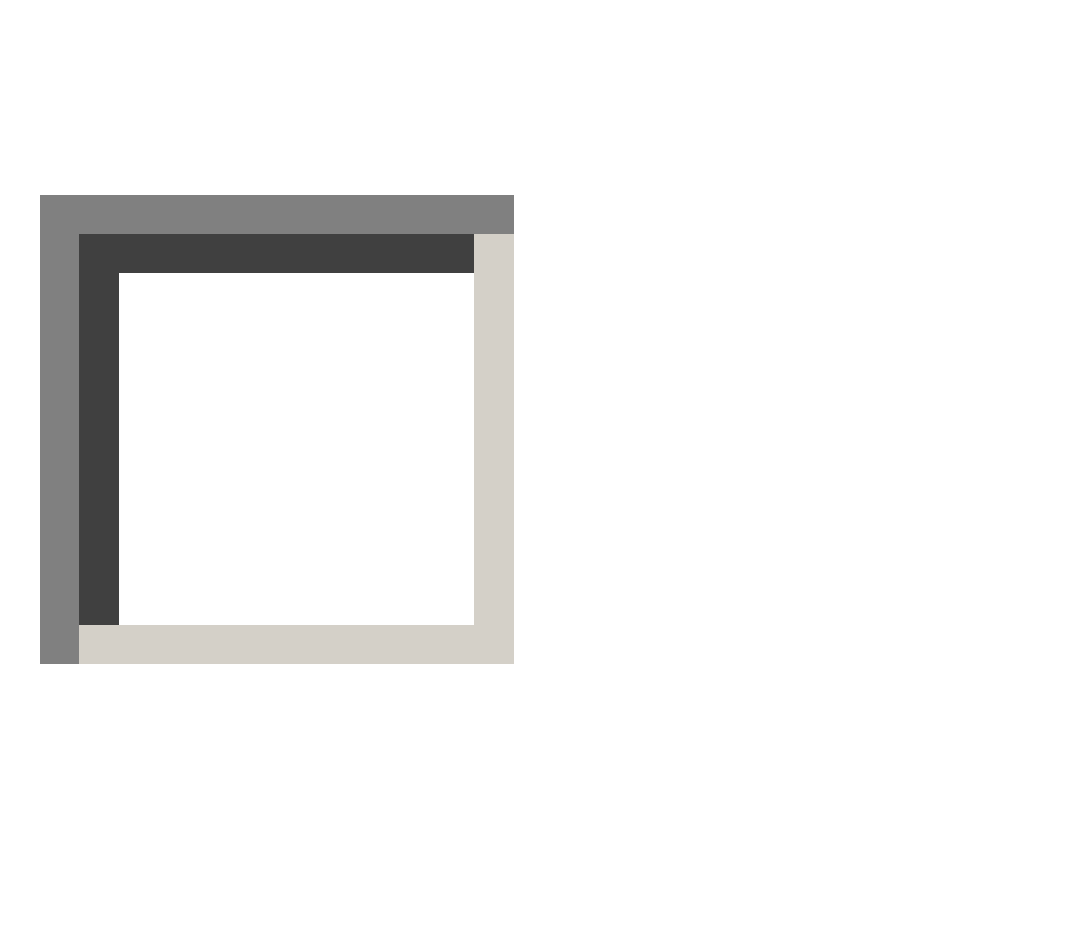
4. Обязательными свойствами хорошего лингвистического корпуса являются

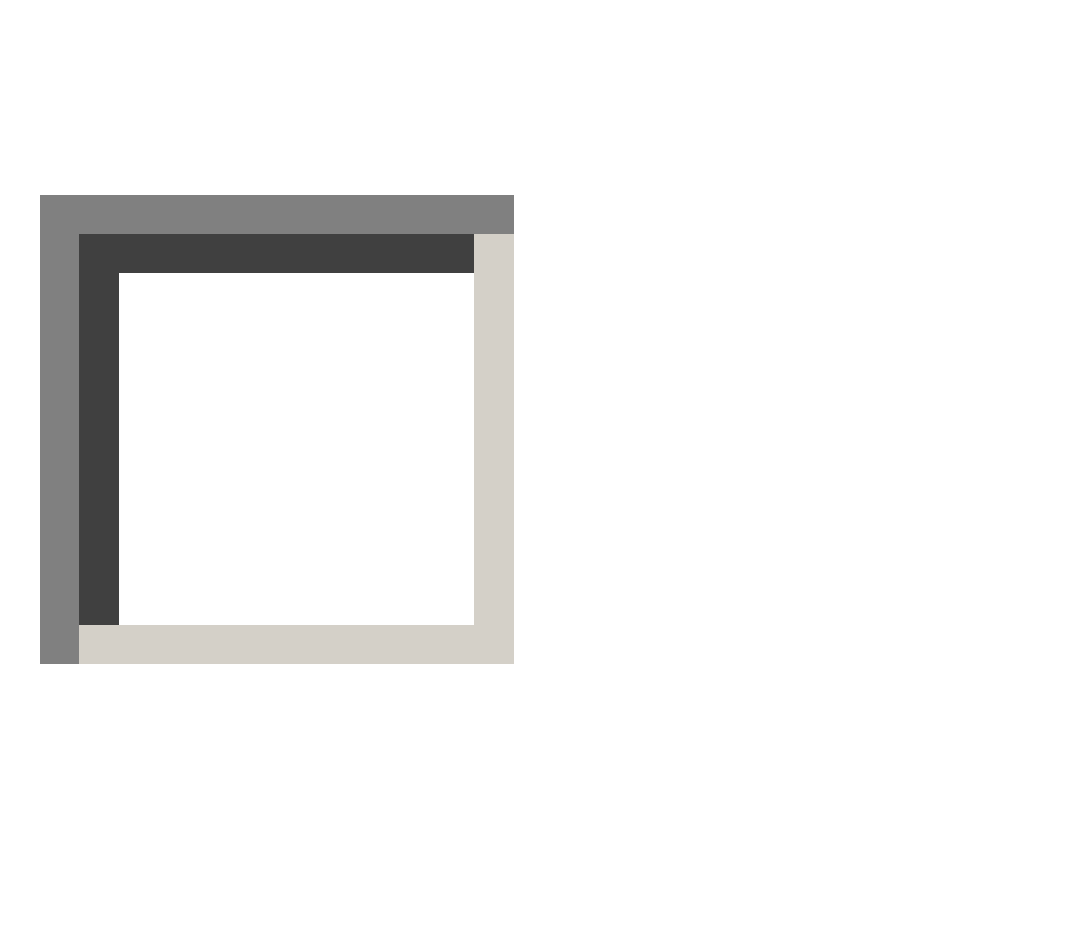
представительность.

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.

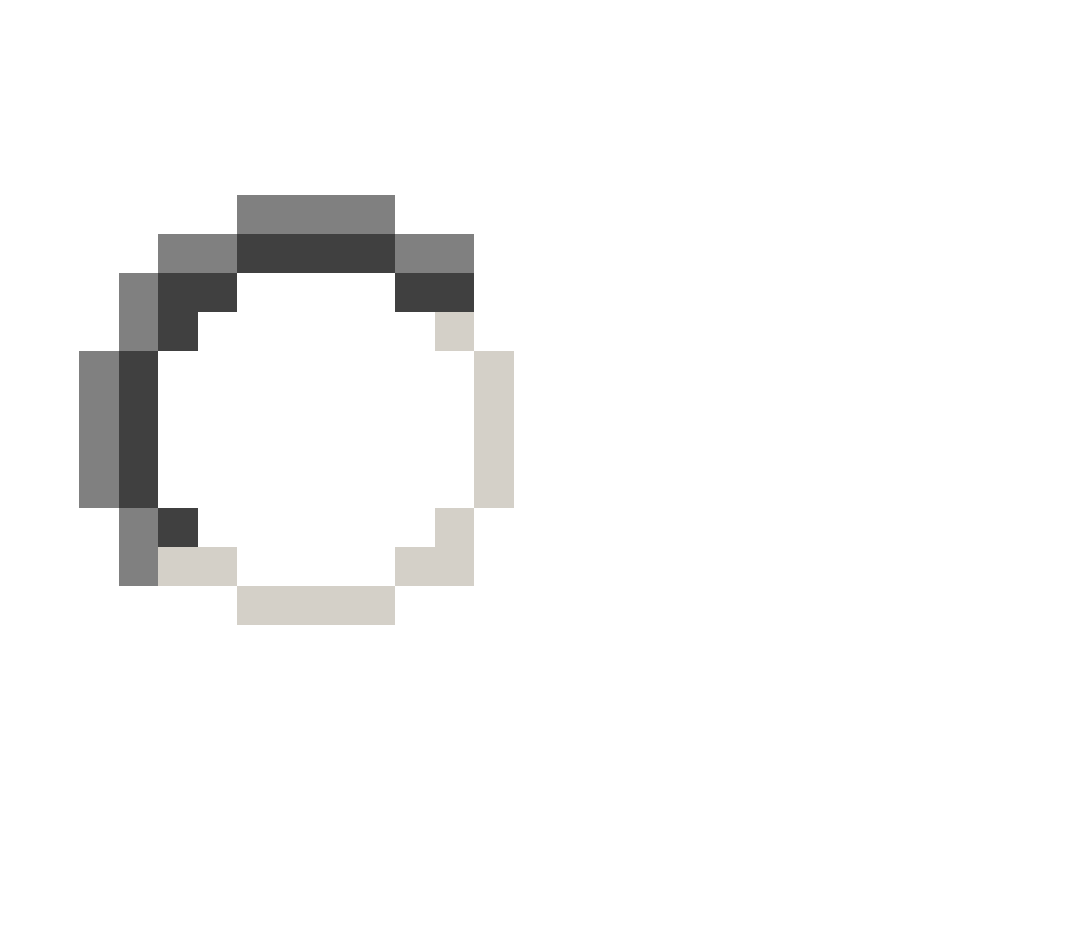
8. Количество найденных результатов обозначается в НКРЯ как количество «словоформ».

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

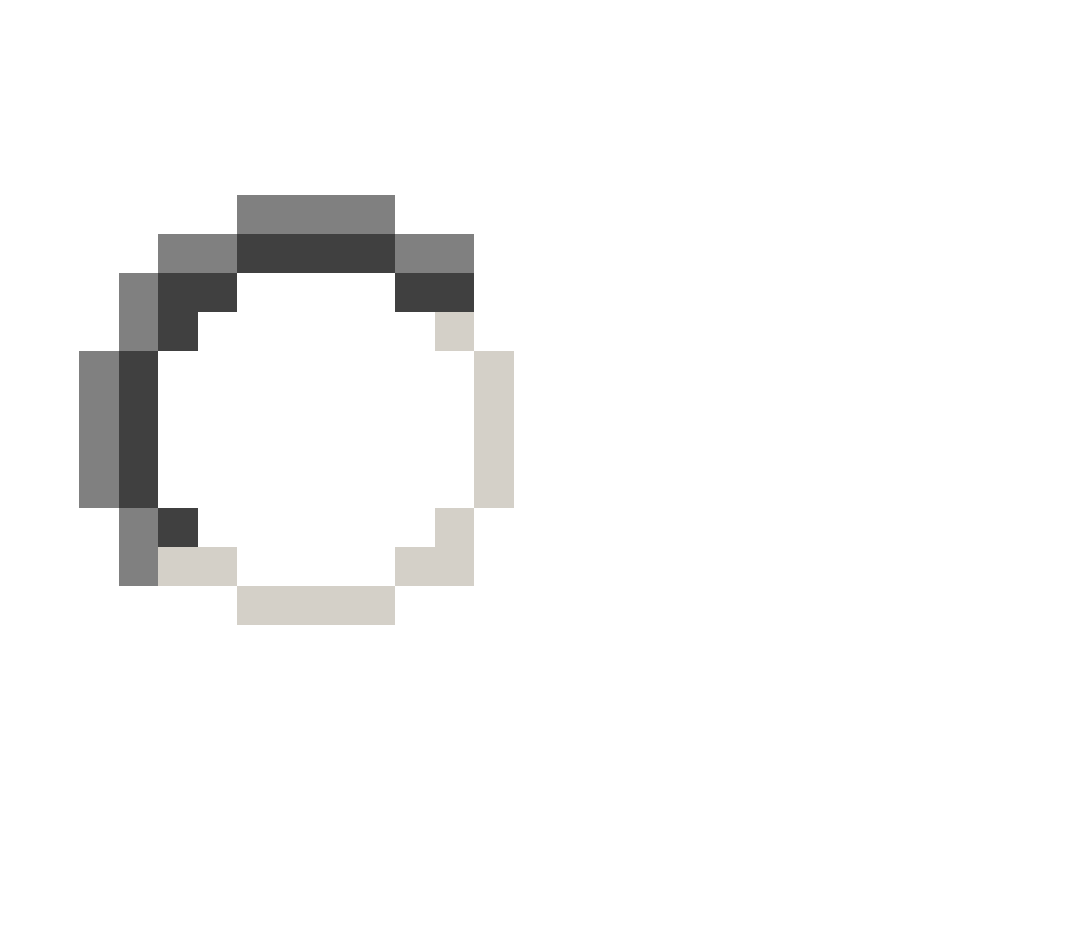
10. Пропорциональное представление в корпусе текстов разных типов называется «репрезентативность» корпуса.

**Выберите единственный правильный вариант ответа.**

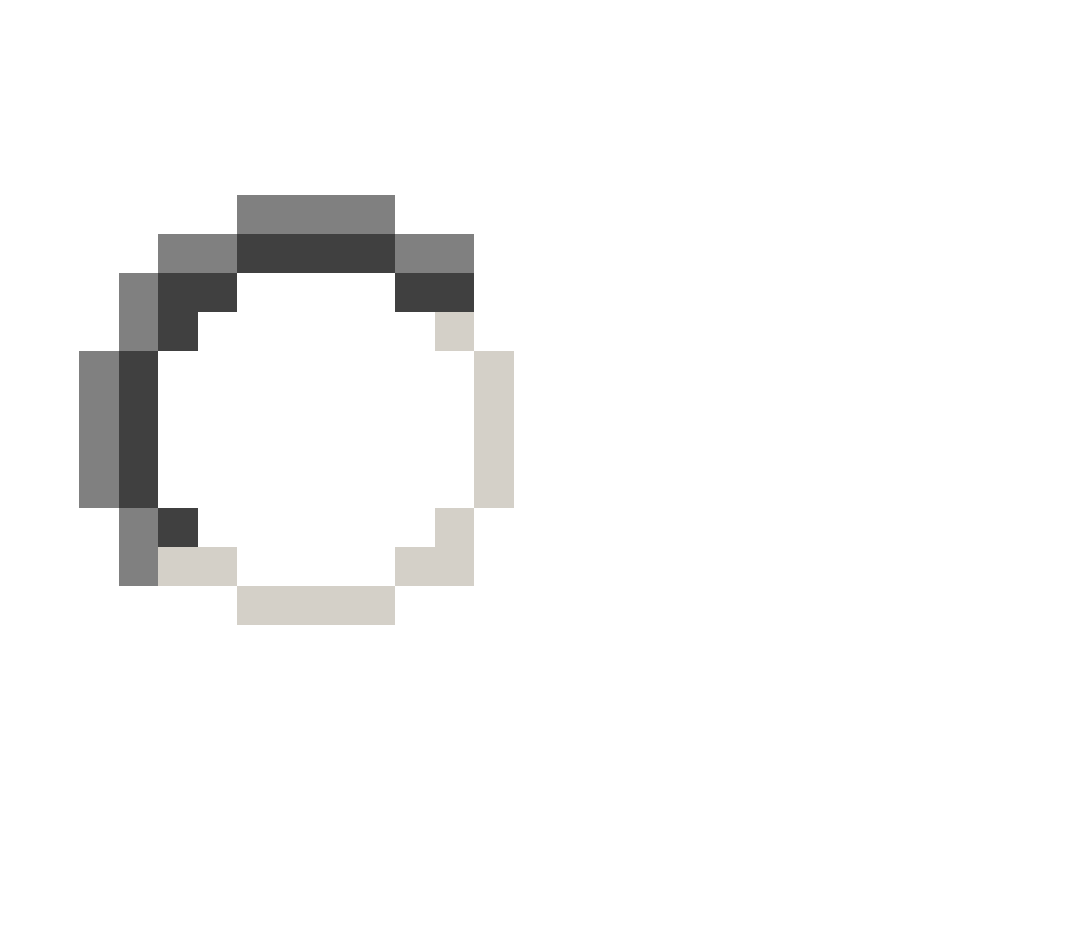
11. Что означает понятие «накопительная сбалансированность»?

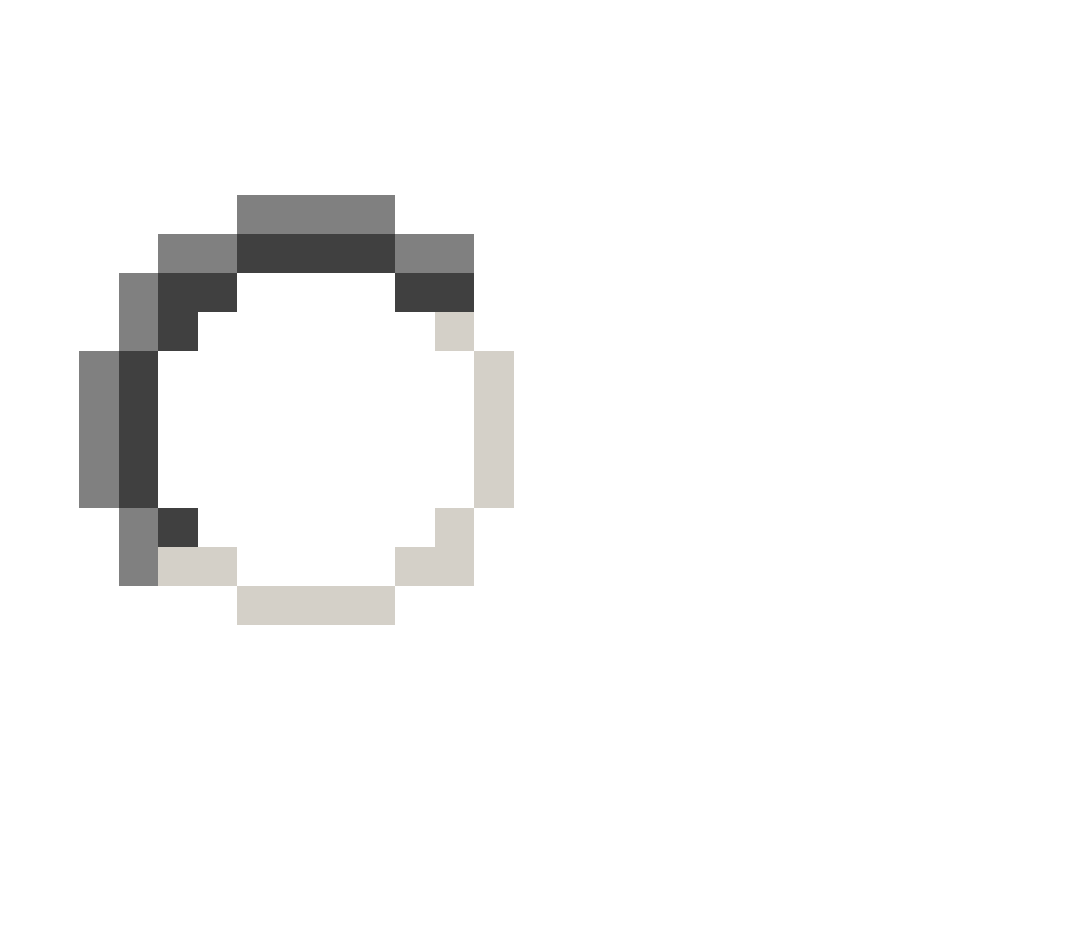
С увеличением числа текстов, входящих в корпус, увеличивается необходимость разнообразить состав включаемых в него документов.

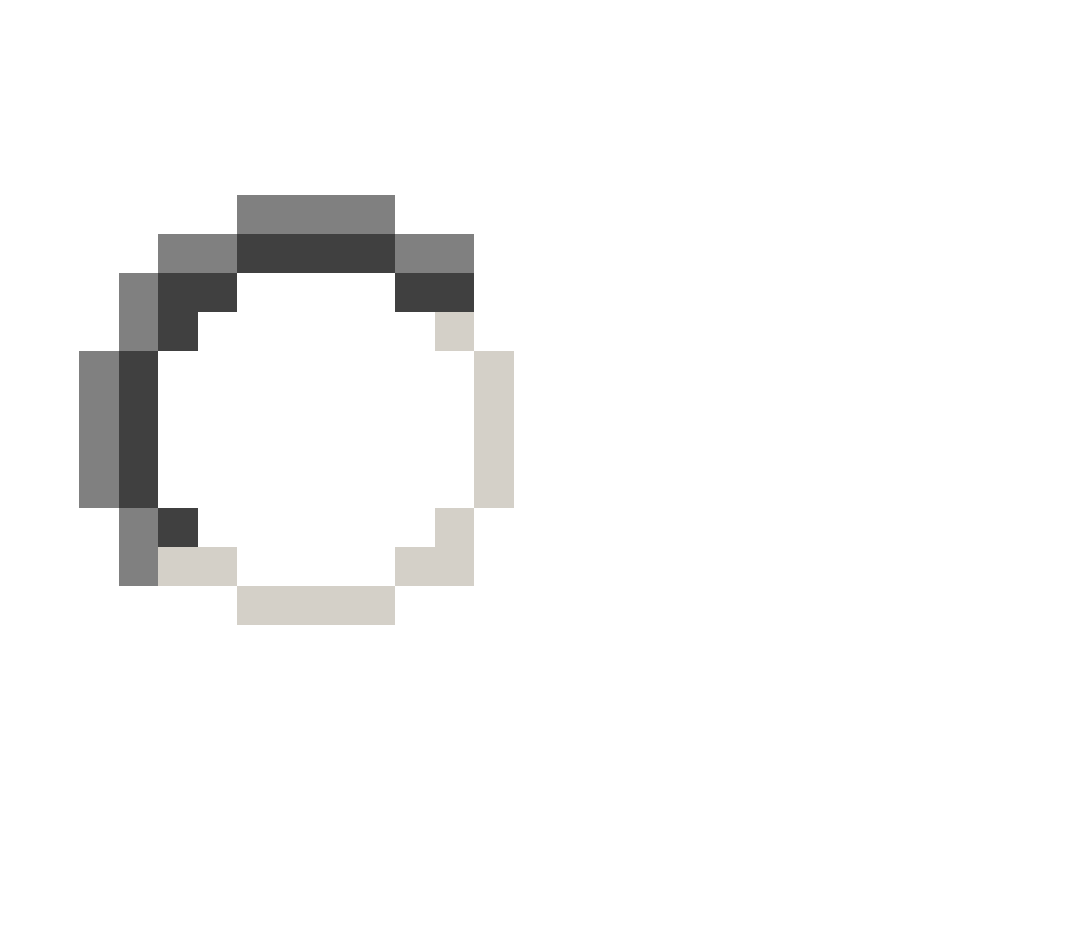
12. Что отличает разметку Обучающего корпуса от аннотации других частей НКРЯ?.

Она проведена в соответствии со школьной грамматикой.

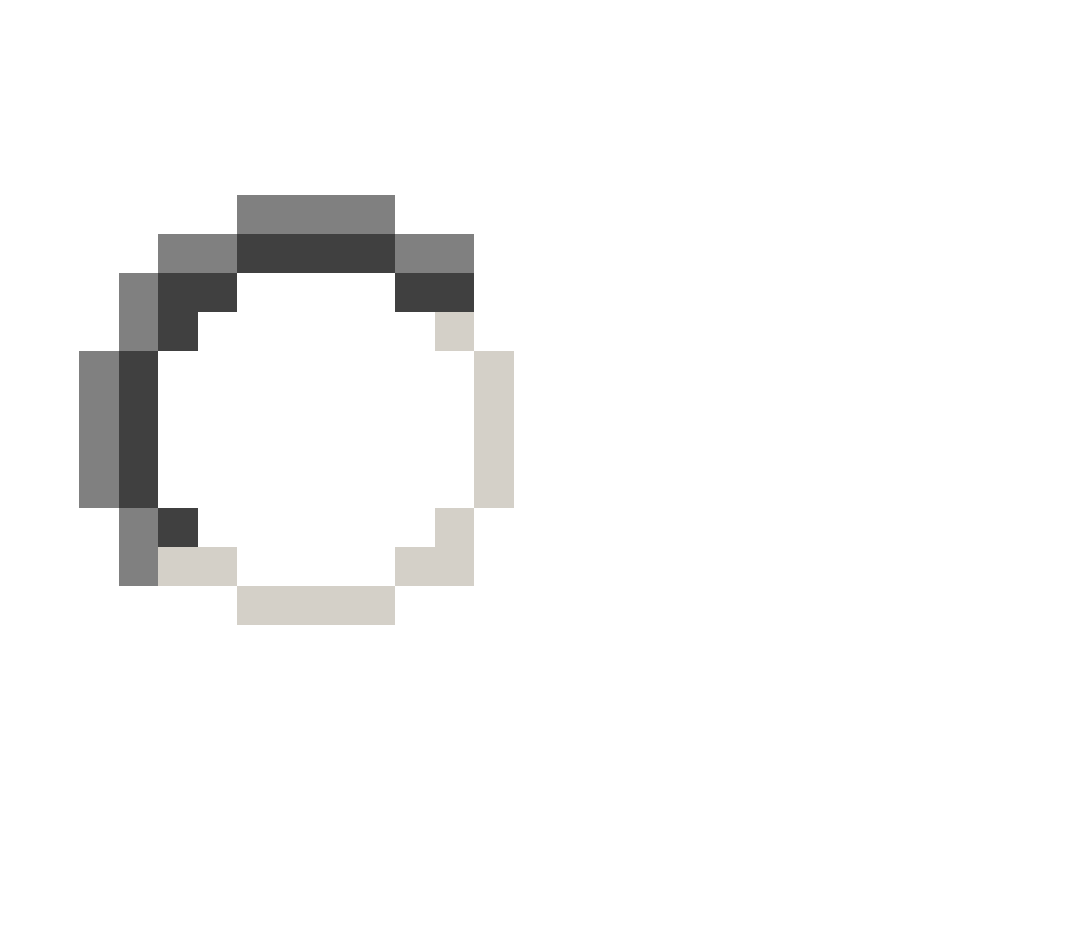
13. Чем отличается лингвистический корпус от электронного словаря?

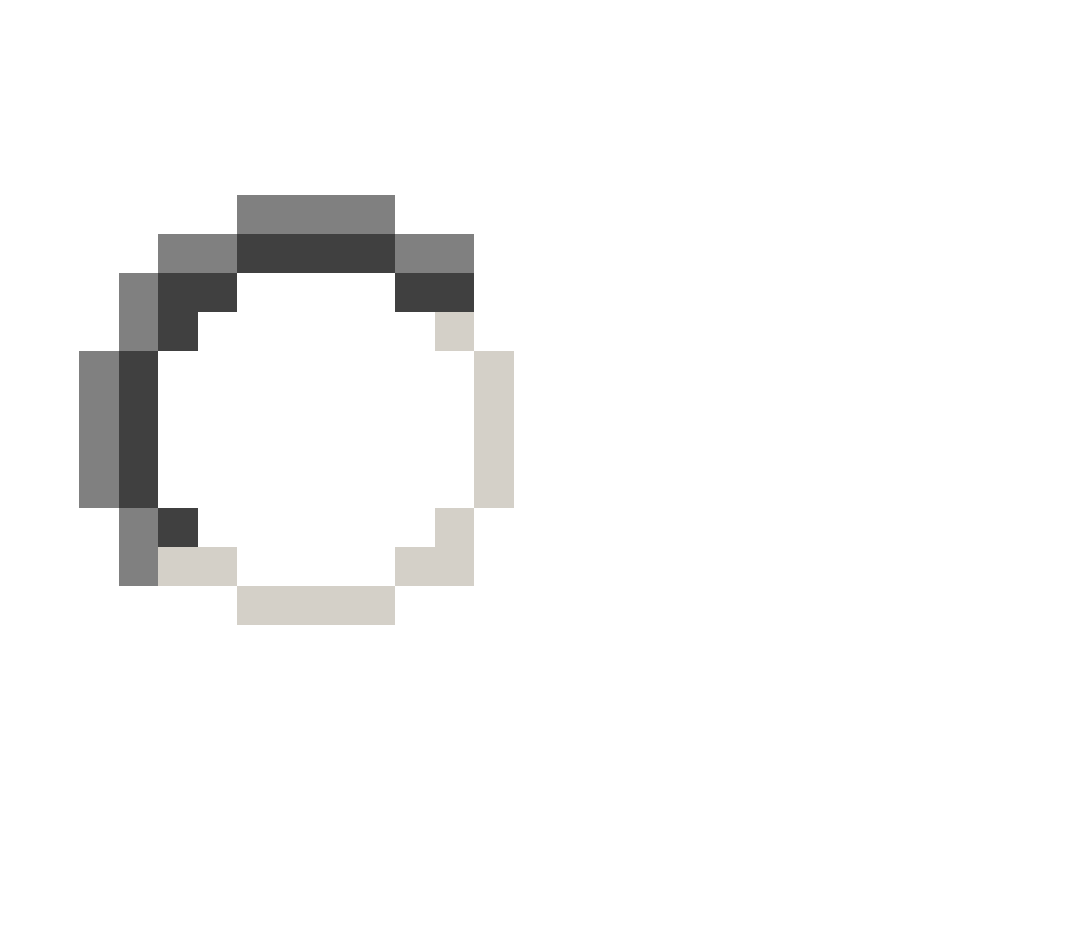
В лингвистическом корпусе языковая норма отражена только опосредованно.

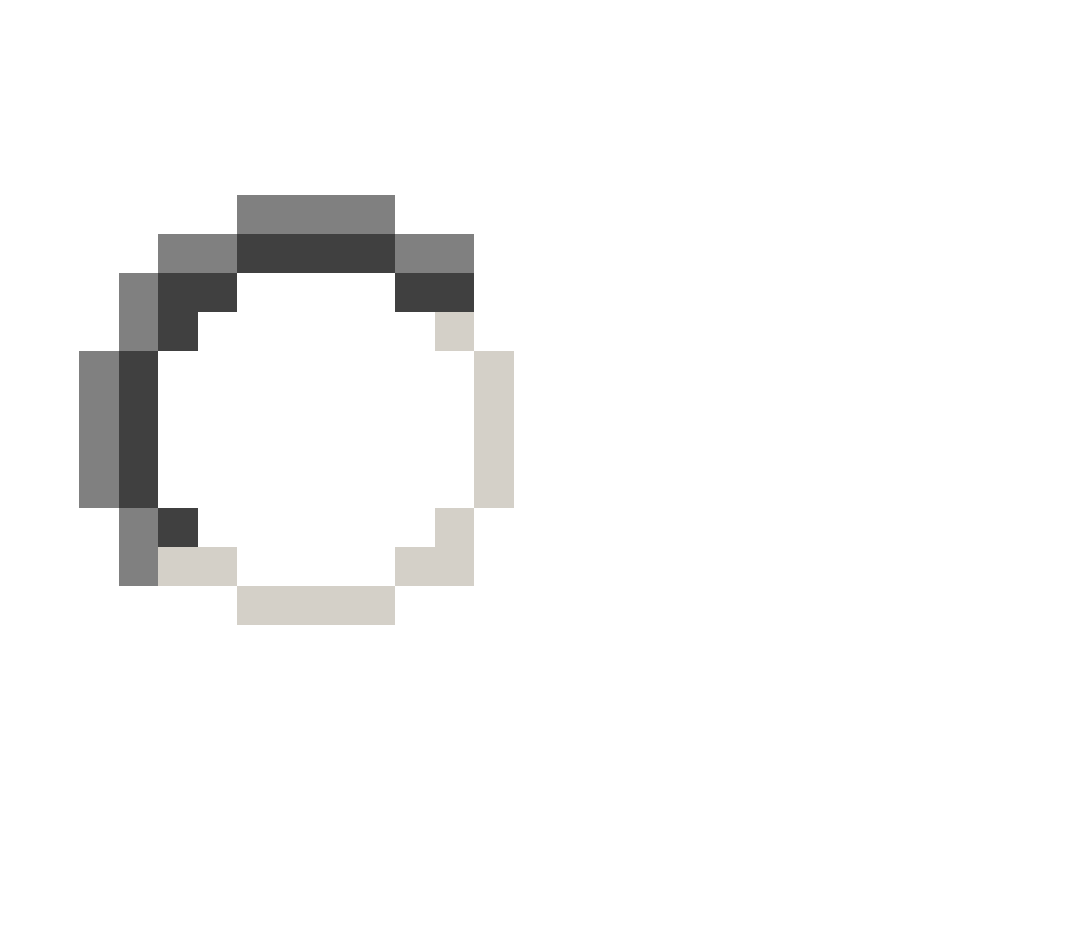
Материалы лингвистического корпуса не дают представления о том, насколько слово распространено в языке.

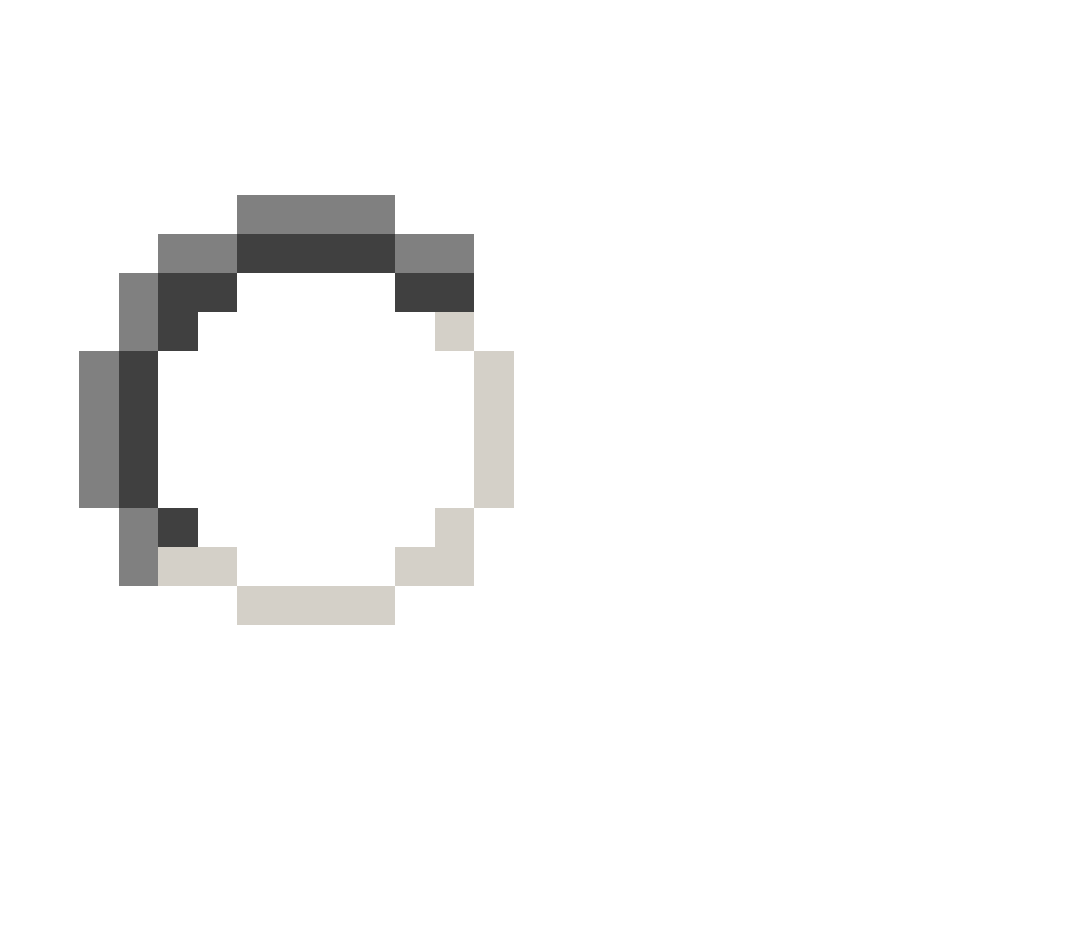
В электронном словаре могут содержаться сведения о семантике слова.

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

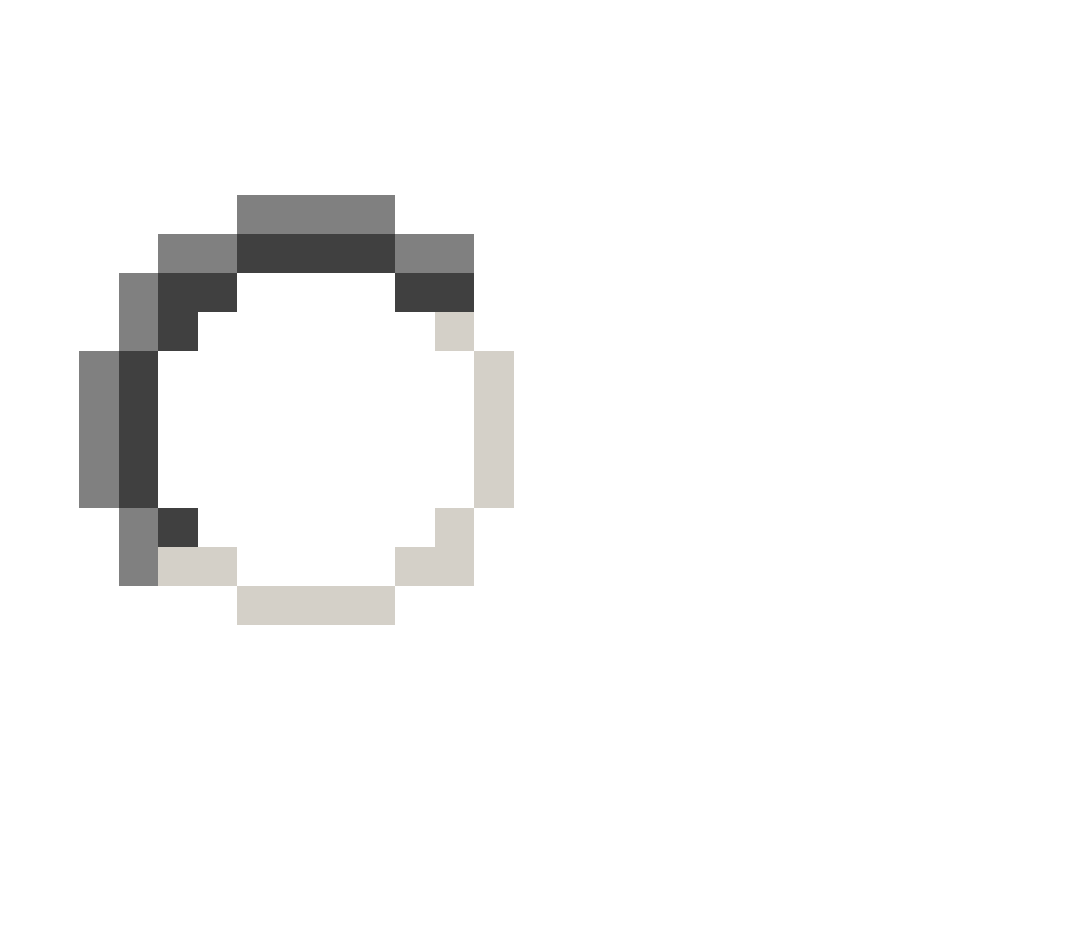
Поиск в пределах одного сайта.

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

Тест”Разметка корпусов”

**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.